Rayyan Systems - Whitepaper

Beyond the Prescription

A Decision Framework for Configuring Deduplication Criteria in Rayyan's Systematic Auto Resolver

---

Prepared by:
Mimi M. Kim, PhD
Head of Scientific Affairs
Health Services Researcher | Systematic Review Methodologist |Industry Executive

Sept 2026

© 2026 Rayyan Systems, Inc. All rights reserved.
sales@rayyan.ai www.Rayyan.ai

A COMPANION RESOURCE TO

Kim MM, Shanaa AM, Hammady H, Aboelnour M, Ayan R. Enhancing Transparency, Auditability, and Reproducibility of Deduplication in Systematic Reviews: Tutorial for the Rayyan Method and Systematic Auto Resolver. Interact J Med Res 2026;15:e94317.

About Rayyan

Rayyan is the most widely adopted systematic review platform in active research use, ranking #1 in new PROSPERO protocol registrations since 2025, with over 1.3 million registered researchers across 30,900 organizations in 190 countries. Rayyan's foundational paper has been cited more than 30,000 times. With over 2.2 million reviews conducted and 3.27 billion screening decisions processed to date, Rayyan operates at a scale that far exceeds the published systematic review literature.

In January 2026 alone, researchers created 47,859 new systematic review projects on Rayyan, surpassing the 43,508 systematic reviews indexed in PubMed across all of 2025. By July 2026, Rayyan had supported 422,228 new review projects year-to-date, more than 16 times the number of systematic reviews indexed in PubMed over the same period (Rayyan internal platform data, 2026; PubMed systematic review publication type filter, retrieved August 7, 2026). Production and publication are distinct stages of evidence synthesis: the published record captures only a fraction of the research in progress. This comparison underscores Rayyan's substantial contribution to the global production of systematic reviews each year. That scale of adoption reflects a research community's trust in a platform built on a clear principle: that the transparency, auditability, and reproducibility that rigorous evidence synthesis demands can be delivered without sacrificing efficiency.

The author of this paper is **Dr. Mimi M. Kim**, PhD, Head of Scientific Affairs at Rayyan Systems, where she bridges deep expertise in systematic review methodology with product development, scientific communications, and strategic partnerships. Dr. Kim is a health services researcher and methodologist with over 60 peer-reviewed publications and experience spanning NIH-funded academic research, commercial biotech R&D, FDA regulatory submissions, and evidence synthesis methodology. She brings to this white paper the perspective of a practicing methodologist who has conducted and evaluated systematic reviews across clinical, academic, and regulatory contexts.

Contents

About Rayyan 1
1. Introduction 3
2. Two Stages, Two Questions 3
3. Why a Single Prescribed Combination Would Undermine the Tool's Value 4
4. Choosing Fields: A Decision Framework 5
4.1 Field by Field Considerations 5
4.2 A Diagnostic Approach 7
5. Normalization: Reversible by Design 8
6. A Different Architecture: Why Direct Comparison Breaks Down 9
7. Conclusion 11
Methodological Note: PubMed Systematic Review Count 12
References 13

> ***Rayyan does not prescribe a single optimal combination of deduplication fields. This paper explains why and provides a complete decision framework to use instead.***

1. Introduction

Researchers evaluating Rayyan's Systematic Auto Resolver, whether for their own systematic review or as part of a comparative evaluation of deduplication tools, frequently ask a direct question: what is the optimal combination of fields to apply? It is a fair question to ask of a tool that offers a choice in the first place. In Rayyan, those choices include Title, Author, Journal, Year, Pages, DOI, Publication Type, and Overall Similarity Percentage as selectable criteria, with the ability to select the preferred search file, and whether to apply normalization. Rayyan does not recommend a single prescribed combination for the very reasons choice is provided. This white paper explains the reasons why and provides further guidance instead.

Kim et al. (2026) already establishes an evidence based, field by field reliability assessment and offers an example iterative pass sequence for Systematic Auto Resolver, described explicitly as illustrative for many review types rather than as a universal answer. This white paper is a companion resource to that publication, not a replacement for it. It expands the field-by-field discussion with additional edge cases, translates that guidance into a decision framework keyed to characteristics a researcher can observe in their own dataset, and explains the architectural reason a single prescribed combination is not something Rayyan can offer without undermining the design that makes Systematic Auto Resolver effective. It closes with a discussion of why comparing Rayyan's output to that of single stage deduplication tools requires care, since the two are not producing the same kind of output.

This paper should be read alongside, not in place of, Kim et al. (2026), which remains the primary citable source for the Rayyan Method and Systematic Auto Resolver, including the step-by-step platform workflow, the documentation checklist for methods reporting, and the advanced settings discussed briefly here.

2. Two Stages, Two Questions

Rayyan's deduplication process separates two methodologically distinct acts: detection and resolution (Kim et al., 2026). Detection is a single, deterministic step. Rayyan's algorithm scans the imported dataset and returns a set of possible duplicates, each carrying a similarity score, using the same detection logic for every user and every dataset. This stage is not configurable by the end user and is not where individual judgment enters the process.

Resolution is the second, distinct stage, and it is here that a research team applies its own criteria, in whatever combination and sequence suits the dataset, to determine which possible duplicates should be removed. Because this stage is user directed and repeatable, a team can apply progressively broader criteria across multiple passes, review the results after each pass, and adjust course before anything is finalized. The order in which passes are applied does not change the final set of resolved duplicates; it only determines which specific pass resolves a given record when its criteria overlap with another pass.

Research teams can sequence their own criteria without concern that a different ordering would produce a different outcome.

This separation matters for how comparative claims about Rayyan should be evaluated. A question such as whether Rayyan captures all the duplicates that another tool identifies is a detection question, and it has a stable, comparable answer: recall on the pool of possible duplicates. A question such as what result Rayyan produces when deduplicating a given dataset is a resolution question, and its answer depends entirely on which criteria the user selects, because resolution is designed to be configured rather than fixed. Treating a resolution outcome as though it were a fixed property of the tool, the way detection performance is, is a common source of confusion when Rayyan is evaluated alongside single stage tools. Section 6 returns to this distinction directly.

3. Why a Single Prescribed Combination Would Undermine the Tool's Value

If Rayyan published one recommended field combination as a standard configuration, it would answer a narrower and less useful question than the one most researchers actually have. A fixed combination can only be genuinely optimal for one profile of dataset: one level of DOI completeness, one mix of source databases, one degree of metadata cleanliness. Kim et al. (2026) makes this explicit in its own language, describing its example pass sequence as illustrative for many review types, not as a universal recommendation, and stating that optimizing criteria selection requires an understanding that no single field is perfectly reliable across all records or databases.

A fixed prescription would also work against the design principle that gives Systematic Auto Resolver its precision advantage over single stage automated tools. Rayyan's two stage architecture exists specifically so that detection can be maximally sensitive, casting a wide net, while resolution can be tailored to the specific reliability profile of a given dataset. Collapsing that flexibility into one fixed combination would mean treating a dataset with clean, DOI complete records the same way as one drawn substantially from gray literature and preprints, or else settling on a hedge broad enough to be safe for both that offers little more direction than the field-by-field guidance already provides on its own.

There is also a validity concern worth stating plainly. If Rayyan prescribed a single combination and a researcher applied it uniformly without regard to their own data, the resulting deduplication outcome would primarily reflect the appropriateness of that prescribed combination for that specific dataset, not a property of Rayyan's underlying detection or resolution capability. A single combination applied uniformly across heterogeneous datasets effectively becomes a test of the combination rather than a test of the tool, which is a validity problem for anyone conducting a comparative evaluation. What can be fairly and consistently evaluated across tools is addressed in Section 6.

4. Choosing Fields: A Decision Framework

This section expands on the field-by-field reliability guidance in Kim et al. (2026) and translates it into a set of diagnostic questions a research team can apply to its own dataset. The goal is not to arrive at a single recommended combination but to help a team reason from the observable characteristics of their data to a defensible starting configuration, consistent with the published literature on deduplication field performance.

4.1 Field by Field Considerations

Title. Title is consistently populated and comparatively stable across databases, but formatting still varies in case, punctuation, and the inclusion of subtitles (Bramer et al., 2016; Kim et al., 2026). Text normalization, which lowercases text and removes punctuation, is recommended whenever Title is used in matching. One edge case worth noting beyond the manuscript's treatment: very short or generic titles, such as single word editorial headers or recurring section titles, can appear across many unrelated records. When Title is combined only with a weak secondary field, this can introduce false positives even after normalization. A related edge case involves special characters and script variants within titles, for example a mathematical symbol such as the Greek letter beta appearing in one database record while another database spells the term out, or a similar-looking but computationally distinct character from a different script being substituted for it. These cases can be resistant to both matching rules and normalization, since the underlying text differs at the character level even when a human reader would recognize the titles as identical. Overall Similarity based detection, described further below, can still surface these as possible duplicates even when an exact-match Title rule cannot, though they will typically require manual resolution. For this reason, Title is best used in combination with Year and Author from the first pass onward rather than in isolation.

Author. Author fields commonly vary in format across databases, including full name versus initials and differences in surname ordering, which can reduce recall when Author is used alone as an identifier (Kim et al., 2026). An additional edge case is author list length: some databases truncate longer author lists (for example, listing the first three authors followed by “et al.”) while others report the full list. Matching on the complete author string can therefore fail even between two records for the same article. Because Systematic Auto Resolver matches on the complete author string rather than an isolated first author, this kind of truncation difference can cause an exact-match Author criterion to fail even between two records for the same article. For datasets where author list formatting is inconsistent across sources, Title and Year are typically a more robust primary pairing, with Author reserved as a confirmatory field once formatting consistency has been verified or supplemented with Overall Similarity to catch near-matches that a strict Author match would otherwise miss.

Journal. Journal fields are comparatively reliable when records are drawn from well indexed biomedical journals, but less so for gray literature, conference proceedings, or preprints (Kim et al., 2026). This is compounded by a documented, recurring source of mismatch: databases frequently export journal titles

differently, some in abbreviated form and others as the full title, which can prevent an otherwise identical record pair from matching on Journal alone (Bramer et al., 2016). Where a dataset merges exports from multiple databases, Journal is best treated as a confirmatory field in a later pass rather than a primary matching field in the first pass.

DOI. DOI is a high precision field. A confirmed DOI match is strong evidence of a true duplicate. Its lower recall stems from documented, real sources of error: duplicate DOIs assigned across records, incorrect DOIs present in source metadata, DOIs that are simply not registered for a given record in a given database, and, less commonly, character recognition errors in DOI transcription (see the DOI error typology reported in matching studies such as the Pure and SciVal case study cited below). Kim et al. (2026) accordingly recommends DOI as a supplementary, confirmatory criterion applied in a later pass rather than as a primary or sole matching field. This is consistent with published practice describing DOI focused deduplication passes as an extension applied after, not instead of, more foundational fields (Main, 2023).

Pages and Publication Type. Both fields are affected by formatting inconsistency across databases, including first page only versus full page range, and online first versus final print pagination (Kim et al., 2026). Bramer et al. (2016) specifically flag page number formatting as a frequent, non-substantive source of apparent difference between records that are otherwise true duplicates. Publication Type carries an additional risk beyond formatting: large scale comparisons of Scopus and Web of Science have found that the databases disagree on the type assigned to a given record at a non-trivial rate, with overall misclassification of approximately 2.7 percent for Scopus and 2.3 percent for Web of Science across more than 27,000 documents, most commonly involving articles and reviews being mistaken for one another (Maisano et al., 2025). A related study found that this disagreement is partly structural: Web of Science permits a record to carry two document types at once, for example pairing article with proceedings paper, while Scopus restricts every record to a single type. Records receiving this kind of dual classification showed a meaningfully higher misclassification rate than singly classified records in both databases (Maisano et al., 2026). For a dataset merging exports from both databases, this means Publication Type disagreement can reflect a genuine difference in classification philosophy rather than a simple data entry inconsistency. These fields are best used as supplementary criteria within a multi field combination rather than as primary matching criteria.

Overall Similarity Percentage. This field is most useful for capturing duplicates with inconsistent metadata spread across multiple fields, where no single field reliably distinguishes true duplicates from unique records. Kim et al. (2026) recommends beginning at a threshold of 95 percent or higher, which approximates exact matching with tolerance for minor variation, and reviewing results before considering a lower threshold. One extension worth noting: for datasets with substantial metadata corruption, for example legacy records, hand entered citations, or records derived from optical character recognition, a lower similarity threshold may be necessary to achieve adequate sensitivity, but this should be paired with a corresponding increase in manual review capacity, since a lower threshold surfaces more borderline cases requiring human judgment.

4.2 A Diagnostic Approach

Rather than asking which combination is best in the abstract, a research team can work through the following questions about its own dataset. Each question routes to a starting configuration and a rationale grounded in the field considerations above. However the questions are answered, the resulting configuration converges on the same underlying pattern established in Kim et al. (2026): begin with high specificity criteria, broaden progressively across passes, and reserve Overall Similarity thresholds as a final layer subject to manual review before confirmation. This diagnostic step is not a prerequisite. Many datasets resolve well with a first pass using reasonable defaults, and the questions below are most useful after that pass, not before it. They become worth the time once a first pass leaves enough rule-resistant duplicates that manual review alone is no longer practical.

Dataset characteristic	What to check	Implication for criteria selection
Source composition	Is the dataset a single database export, or merged from multiple databases and sources?	Merged, multi-source datasets carry higher risk of Journal and Pages formatting inconsistency. Deprioritize these fields in early passes and enable text normalization before any Title based pass.
DOI coverage	Spot check a sample of 50 to 100 records: is DOI populated and well-formed across most records?	High, consistent coverage supports DOI as a confirmatory field in a later pass. Low or inconsistent coverage means DOI should be deferred to a narrow confirmatory pass or omitted rather than used as a primary criterion.
Literature type	Does the dataset include gray literature, conference proceedings, or preprints alongside indexed journal articles?	Where gray literature is present, Journal field reliability drops further. Weight Title and Author more heavily and consider introducing Overall Similarity earlier as a safety net rather than reserving it for a final pass only . Include Publication Type so that a preprint and its version of record are treated as distinct versions of a publication rather than as interchangeable representations of it, and leave that multiplicity to manual resolution rather than automatic collapse.
Metadata and script complexity	Does the dataset include non-Latin script records, heavily abbreviated metadata, or legacy or OCR derived records?	Verify that text normalization performs as expected on a representative sample before relying on it at scale. Consider a lower Overall Similarity threshold paired with increased manual review capacity.
Review stakes and reviewer capacity	Is this a regulatory or high-stakes review with reviewer capacity for close manual review, or a rapid review with limited capacity?	Higher stakes and greater capacity favor more, narrower passes with manual review after each. Lower capacity may call for fewer, broader passes, but Systematic Auto Resolver confirmations remain irreversible regardless, so verification before confirming

Dataset characteristic	What to check	Implication for criteria selection
		each pass still applies. Deleted records can be recovered through a separate re-import process if needed, but this is not a substitute for reviewing results carefully before confirming a pass.

Because these diagnostic questions are answered from the researcher's own data rather than from a fixed rule, two teams working with genuinely different datasets can reach different, equally defensible starting configurations. That is the intended outcome. Omitting Publication Type from criteria has a specific consequence worth flagging: without it, multiplicity is treated as duplicate representations of the same publication, when it may in fact be different versions, such as a preprint and its later version of record, that a review team would want to evaluate deliberately rather than collapse automatically. It is also why no single configuration in this table, or in Kim et al. (2026), should be read as the recommended combination independent of the dataset to which it is being applied.

5. Normalization: Reversible by Design

Text normalization, converting text to lowercase and removing punctuation before comparison, is one of the more consequential settings available in deduplication tooling, and different architectural choices about when and how normalization is applied carry meaningfully different tradeoffs.

Consider a hypothetical platform in which normalization is applied before duplicate detection, as a fixed step early in the pipeline rather than as a configurable, optional pass. In that architecture, detection operates against a standardized representation of the data rather than the original record as reported. If the normalization rules suit the dataset well, detection sensitivity improves. If the rules are poorly suited to the dataset, for example if they strip information that was in fact meaningfully distinguishing between two records, detection sensitivity is degraded in a way that is difficult to inspect and cannot be reversed within that same workflow, because there is no path back to comparing against the unnormalized record once the pipeline has run.

Rayyan's approach differs in kind. Normalization is an optional advanced setting applied within a given Systematic Auto Resolver pass (Kim et al., 2026). A research team can enable it for one pass and disable it for another, and the underlying record data is never altered, only the comparison performed against it. If a team suspects normalization is producing unexpected results for a particular dataset, for example one containing non-Latin script or extensively abbreviated metadata, they can disable it, inspect the outcome, and re-enable it once satisfied, without losing access to the original record at any point. Kim et al. (2026) recommends exactly this kind of verification for such datasets before proceeding with full scale resolution.

This reversibility is a direct consequence of the two-stage architecture discussed in Section 2. Because resolution is separated from detection and is itself repeatable, individual settings within resolution, including normalization, inherit that same repeatability. A normalize-first architecture, by contrast, folds this decision into the detection stage itself, where it cannot be revisited without re-running the entire process from the original data.

Normalization rules are themselves generalizations, and their safety depends on the characteristics of the dataset rather than being uniformly safer or riskier in the abstract. Consider a rule that standardizes page number formatting: many articles legitimately share the same starting page number, so aggressive page normalization can inflate apparent matches rather than resolve genuine ones, which is the opposite of the failure mode normalization is usually meant to prevent. The special-character and script-variant edge case discussed in Section 4.1 illustrates the same tradeoff from the other direction, where stripping characters can cause a true match to be missed rather than a false one to be created. Taken together, these examples show that more aggressive normalization can resolve more true duplicates or introduce more false positives, depending entirely on whether the rule's assumptions fit the dataset it is applied to.

6. A Different Architecture: Why Direct Comparison Breaks Down

Deduplication tools in the systematic review space take several different architectural approaches, each with its own strengths and limitations. Understanding these patterns, described here in general terms rather than attributed to any specific named platform, clarifies why a single benchmark, such as final duplicate count or a single accuracy percentage, does not capture what is actually being compared when Rayyan is evaluated alongside other tools.

Pattern one: a fixed, single pass automated algorithm. Consider a platform that applies one fixed algorithm to produce a single determination, duplicate or unique, for each record. Its strength is consistency within a single configuration: given the same matching rules, the algorithm applies them the same way every time. However, this consistency has a noteworthy limitation. For architectures that resolve duplicates in-line during import and default to keeping whichever record was imported first, the outcome can still depend on the order search files are uploaded, since a preprint imported before its version of record could be the copy retained, with a different outcome possible had the files been imported in a different sequence. Its limitation is inflexibility: the matching logic embedded in that algorithm cannot be adjusted by the end user for a dataset whose reliability profile differs from what the algorithm was tuned for, for example a dataset with unusually low DOI coverage or heavy representation of gray literature, without a vendor side change to the underlying tool.

Pattern two: a manual, stepwise protocol. Consider a platform where a researcher executes a documented sequence of manual steps directly within a reference manager interface, applying different field combinations at each step. Its strength is transparency: a trained user understands exactly what each step is doing and why. Its limitation is that the steps, once defined, are typically applied once in a

fixed sequence for a given dataset. There is no separate, reusable resolution stage that can be reconfigured and reapplied to the same detected candidate pool without repeating the underlying process, which makes it harder to iterate quickly if an early step proves too aggressive or too conservative.

Pattern three: a normalization first automated pipeline. As discussed in Section 5, consider a platform where text normalization is a fixed, non-optional step that occurs before detection. Its strength is that, when the normalization rules suit the dataset, detection sensitivity can improve meaningfully. Its limitation is that the normalization step is not inspectable or reversible within the workflow, so a poor fit between the normalization rules and a particular dataset degrades results in a way that is difficult to detect after the fact. It is also worth noting that a normalize-first pipeline typically involves three sequential steps, import, normalize, then detect-and-resolve, even where it is perceived and marketed as a single automated stage. Correction under this model is retrospective: a reviewer can flag that a result was wrong after the fact but cannot direct the process before it runs. Rayyan's two-stage design collapses this into fewer, user-directed steps, with normalization available as an optional property of a resolution pass rather than a required stage a reviewer has no way to instruct in advance.

Pattern four: a single similarity threshold tool. Consider a platform that reduces matching to one overall similarity score with a single adjustable threshold, without allowing individual fields to be weighted or excluded. Its strength is simplicity: one setting to configure, one number to interpret. Its limitation is that it cannot distinguish between fields of genuinely different reliability, a well indexed Journal field and a poorly formatted Pages field contribute to the same aggregate score with no ability to discount the less trustworthy one, which can obscure the source of false positives or false negatives when they occur.

Rayyan's two stage architecture is intended to combine the strength of pattern one, a consistent, non-configurable detection stage benchmarkable in its own right, with a resolution stage that avoids the limitations of patterns two through four: it is reusable and repeatable rather than a one-time sequence, it treats normalization as optional and reversible rather than fixed, and it allows individual fields to be selected, weighted through combination, and excluded based on their known reliability for a given dataset rather than collapsing everything into one score. Rayyan also addresses the import-order question directly rather than leaving it to upload sequence: a research team can select which imported source to default to for a given resolution pass; for example, prioritizing a source with more complete DOI coverage or more reliable metadata for downstream screening, rather than having that choice made implicitly by the order files happened to be uploaded.

> ***What's fairly comparable across every architecture is detection recall. What isn't is a single resolution "winner" — because resolution in Rayyan is intentionally configurable, not fixed.***

The practical implication for anyone conducting a comparative evaluation is this: a study measuring final duplicate count or a single accuracy figure across platforms is implicitly treating these different kinds of output as equivalent. A fixed algorithm's output is one answer, arrived at the same way every time. Rayyan's output at any given point in the resolution process is one of many possible answers, contingent on which criteria the research team has chosen to apply so far. What is directly and fairly comparable across all of these architectures is detection performance, meaning recall on the pool of possible duplicates prior to any resolution decision, since this is the one stage common to every approach and not subject to user configuration in any of them.

7. Conclusion

Rayyan does not prescribe a single optimal combination of deduplication fields, and this white paper has aimed to explain both why that is the case and what a research team should do instead. The reasoning is not a limitation dressed up as a feature. A fixed prescription would only be genuinely optimal for one dataset profile. It would also work against the two-stage architecture that gives Systematic Auto Resolver its precision advantage, and it would introduce a potential validity problem for any comparative evaluation that applied it uniformly across heterogeneous data.

In place of a fixed prescription, this paper has offered an expanded field by field reliability discussion, a diagnostic framework keyed to observable dataset characteristics, and an explanation of why direct output comparisons across differently architected deduplication tools require care. Together with the field guidance, example pass sequence, and documentation checklist already published in Kim et al. (2026), this is intended to function as a complete, self-sufficient resource for a research team configuring Systematic Auto Resolver for their own dataset, and for an evaluator designing a fair, methodologically sound comparison across deduplication tools.

This white paper is a companion resource to, and should be cited alongside, Kim et al. (2026) as the primary source for the Rayyan Method and Systematic Auto Resolver.

Methodological Note: PubMed Systematic Review Count

PubMed systematic review counts reported in this white paper were retrieved using the publication type filter systematic review[pt], applied via the PubMed advanced search interface at pubmed.ncbi.nlm.nih.gov. This filter identifies records formally designated as systematic reviews in the MEDLINE indexing process, and represents a conservative, independently reproducible count of published systematic reviews.

Two searches were conducted and retrieved on August 7, 2026:

- Search 1 (2025 full year): systematic review[pt] filtered to January 1, 2025 through December 31, 2025. Result: 43,508 records.
- Search 2 (2026 year-to-date): systematic review[pt] filtered to January 1, 2026 through August 7, 2026. Result: 24,893 records.

The 2026 year-to-date figure should be interpreted with the understanding that PubMed indexing for recently published articles is ongoing. Citations submitted for indexing within the preceding six to eight weeks may not yet appear in search results, meaning the 2026 figure will increase as indexing is completed. This indexing lag is a standard characteristic of the PubMed database and does not affect the 2025 full-year figure, which reflects a fully indexed and settled record. Both searches used an identical methodology to ensure comparability across time periods.

Rayyan new review project counts reflect internal platform data for the period January 1 through July 31, 2026, defined as new systematic review projects created by registered users within the Rayyan platform during the specified period.

References

1. 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. doi:10.3163/1536-5050.104.3.014
2. Jiang Y, Lin C, Meng W, Yu C, Cohen AM, Smalheiser NR. Rule-based deduplication of article records from bibliographic databases. Database. 2014;2014:bat086. doi:10.1093/database/bat086
3. Kim MM, Shanaa AM, Hammady H, Aboelnour M, Ayan R. Enhancing Transparency, Auditability, and Reproducibility of Deduplication in Systematic Reviews: Tutorial for the Rayyan Method and Systematic Auto Resolver. Interact J Med Res. 2026;15:e94317. doi:10.2196/94317
4. Madsen HH, Madsen D, Gauffriau M. Evaluation of unique identifiers used as keys to match identical publications in Pure and SciVal, a case study from health science. F1000Res. 2016;5:1539. doi:10.12688/f1000research.8913.2
5. Maisano DA, Mastrogiacomo L, Ferrara L, Franceschini F. A large scale semi automated approach for assessing document type classification errors in bibliometric databases. Scientometrics. 2025;130:1901-1938. doi:10.1007/s11192-025-05244-y
6. Maisano DA, Ferrara L, Franceschini F. Single versus dual document type assignment: an empirical comparison of classification policies in Scopus and Web of Science databases. Scientometrics. 2026. doi:10.1007/s11192-026-05602-4
7. Main E. Extending the Bramer Method with EndNote 20: Deduplicating with DOI. J Can Health Libr Assoc. 2023;44(2):48-63. doi:10.29173/jchla29720
8. McKeown S, Mir ZM. Considerations for conducting systematic reviews: evaluating the performance of different methods for de-duplicating references. Syst Rev. 2021;10(1):38. doi:10.1186/s13643-021-01583-y