https://doi.org/10.51408/issi2025_110

How systematic are the systematic reviews?

Andrey Guskov[1], Denis Kosyakov[2], Irina Selivanova[3], Alexandra Malysheva[4]

[1]*guskov.andrey@gmail.com*
Russian Centre for Scientific Information, Leninsky pr., 32A, Moscow (Russia)
Institute of Computational Mathematics and Mathematical Geophysics SB RAS,
Ac. Lavrentieva ave. 6, Novosibirsk (Russia)
Russian Institute of Economics, Policy and Law, Dobrolubova Str. 20A, Moscow (Russia)

[2]*kosyakov@sciencepulse.com*
Institute of Computational Mathematics and Mathematical Geophysics SB RAS,
Ac. Lavrentieva ave. 6, Novosibirsk (Russia)
Russian Institute of Economics, Policy and Law, Dobrolubova Str. 20A, Moscow (Russia)

[3]*i-seli@yandex.ru*, [4]*bag_bala@mail.ru*
Russian Institute of Economics, Policy and Law, Dobrolubova Str. 20A, Moscow (Russia)

Abstract

Systematic literature reviews (SLRs) are widely recognized as a cornerstone of evidence-based research, providing comprehensive syntheses of existing literature on specific topics. Despite the availability of standardized protocols (e.g., PRISMA), many authors do not fully adhere to established methodological requirements. This study aims to determine how frequently four basic criteria – explicit search strategies, inclusion/exclusion criteria, a complete list of included sources, and a clear model of analysis – are met in publications that are labeled as SLRs.

Using Scopus, we sampled 1000 publications in four disciplines (Medicine, Computer Science, Social Sciences, and Biochemistry) and used large language models to assess compliance with each criterion. Results show that 53% of SLRs satisfy all four requirements, while 16% fail at least two. Search and inclusion criteria are widely recognized as core components of SLRs, while fewer authors provide a complete reference list or adopt an explicit analysis model. Disciplinary differences emerged, with Biochemistry and Medicine having the highest rates of full compliance, and Computer Science the lowest. In Medicine, high-impact journals had a 13% higher compliance rate, demonstrating the impact of journal policies. However, overall compliance did not correlate with citation impact. The prevalence of PRISMA in Medicine and Biochemistry likely drives higher compliance in these fields. Future research will expand the analysis by incorporating additional criteria and expert assessments, providing deeper insight into the role of SLR methodologies and the accuracy of evaluations based on AI-tools.

Introduction

Systematic literature reviews are considered to be one of the main tools of scientific methodology, as they summarize and critically analyze all available literature on a particular topic, forming a reliable evidence base for further research (Mathew, 2022). One of the most important principles of SLR is considered to be comprehensive sourcing, which promotes unbiased conclusions and reduces the risk of missing relevant data (Cooper et al., 2018), which can lead to biased effect estimates and unreliable conclusions (Tricco et al., 2008).

Despite the importance of methodological rigor and the availability of the well-known PRISMA family of protocols, many authors do not always adhere to these

requirements. For example, Norling et al (2023) showed that a large proportion of urology reviews did not report detailed search strategies. A further problem is the lack of detail in the description of inclusion and exclusion criteria: although authors often mention such criteria, the actual details of their application remain unclear (Budgen et al., 2018). Frost et al. (2022) also found that only 8% of protocols met all PRISMA-P requirements, indicating the formal nature of adherence to established methodological standards. Finally, many reviews ignore the recommendation to publish a full list of included sources (Kitchenham et al., 2022) and limit themselves to a general description. As a result, it is not uncommon for reviews that claim to be 'systematic' to actually have a very superficial methodology, while some 'mapping studies' are closer to full-fledged SLRs (Budgen et al., 2018).

Large Language Models (LLMs) are increasingly being used to process the growing amount of scientific information. There are already examples of their successful use to automate the processes of selection, extraction, judgment, analysis and narration in the preparation of SLR, which show results comparable to those of experts (Hasan et al., 2024). However, it remains an open question to what extent review authors themselves correctly specify and apply the underlying methodological principles when assessing the quality of such reviews against the key criteria of transparency and reproducibility. In particular, Budgen et al. (2018) showed that review authors do not always fully and transparently describe the sourcing, inclusion/exclusion, list of selected primary studies, and data analysis model, even though these aspects directly affect the reproducibility of reviews and provide a basis for assessing their methodological quality. However, systematic peer review of these requirements is laborious, making it difficult to regularly analyze the quality of SLRs.

The **aim** of the study is to test, using large language models, how often basic requirements are met in SLRs that are labeled as systematic:

- **R1:** presence of explicitly stated criteria for finding sources,
- **R2:** presence of explicitly stated criteria for inclusion/exclusion of sources,
- **R3:** presence of a list of sources selected for review,
- **R4:** presence of a model for the analysis of sources.

Based on this objective, the **following research questions** are formulated:

RQ1. For what part of SLRs are requirements R1-R4 fulfilled?

RQ2. Are there statistically significant differences in compliance between disciplines?

RQ3. Are these requirements more often fulfilled in high-impact journals?

RQ4. Is there a relationship between completing requirements and citing SLRs?

Method

Four scientific fields were selected for the study in which SLRs have a significant representation (ASJC code in parentheses):

- **Medicine** (2700) has the longest tradition of standardized systematic reviews, particularly under the PRISMA guidelines, and exhibits clear protocols for risk of bias assessment and data synthesis.
- **Computer Science** (1700) has experienced a rapid increase in the number of SLRs, often adapting methodologies from other fields or employing

alternative frameworks such as Kitchenham's guidelines, thus illustrating a discipline in the midst of methodological standardization.

- **Social Sciences** (3300) represent a broad, interdisciplinary arena where systematic reviews are also undertaken but are typically governed by more flexible or mixed-method approaches, providing a contrast to the highly codified medical SLR protocols.
- **Biochemistry** (1300) typifies a natural science discipline that frequently employs SLRs to summarize experimental evidence; it also increasingly intersects with data-driven analyses, making it pertinent for assessing how LLMs handle specialized literature.

In each of these areas, a sample of publications was generated from Scopus that met the following criteria:

- Title or abstract contains "systematic review" OR "systematic literature review",
- Publication year 2022,
- Document type 'Article', 'Review', or 'Conference Paper',
- Open access (any).

From each sample, 400 publications were randomly selected. These were pre-filtered using LLM gpt-o1-mini: the title and abstract were checked to ensure that they were indeed systematic reviews in the specified scientific field. For those that passed, the full text of the publications was downloaded. The text layer was extracted from the PDFs and the number of tokens was calculated (model cl100k of the Python library tiktoken). Publications that appeared to have less than 2,000 or more than 50,000 tokens were discarded. From the publications that passed all checks, 250 were randomly selected for each discipline and a final sample (N=1000) was drawn.

For each article in this sample, the gpt-4o language model was used to determine whether R1-R4 requirements were met, as well as mentions of SLR preparation techniques. Sampling of the results by the article authors showed a satisfactory result.

Results

RQ1. For what part of systematic reviews are requirements R1-R4 fulfilled?

All four requirements are met in 53% of the reviews and 16% of the reviews, which the authors call systematic, do not meet 2 or more requirements (Figure 1). The requirements to specify criteria for finding publications (R1, 89%) and to include them in the review (R2, 93%) are most frequently fulfilled. This is not surprising, since in many journals the requirement to specify where and how publications were searched for and according to which principles they were selected has already become the "gold standard" for SLRs, regardless of the field.

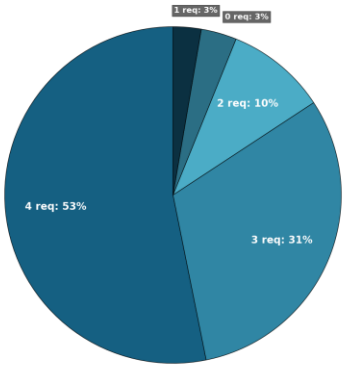

Figure 1. Distribution of SRLs by number of requirements met.

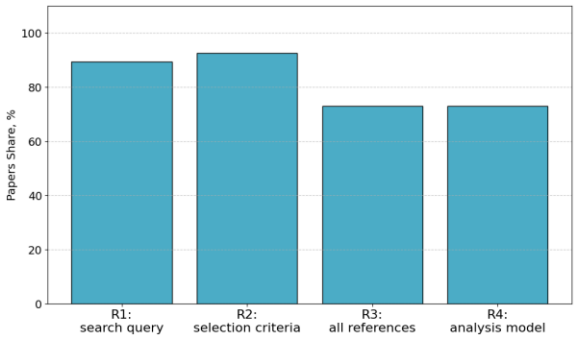

Figure 2. Degree of fulfillment of requirements, entire sample.

The other two requirements 'all references' and 'analysis model' are less frequently fulfilled - only in 73% of the cases each (Figure 2). Moreover, if we consider only 13% of the publications that did not meet exactly two requirements, the majority of them fall, as expected, on this pair (53 out of 93 cases). Most likely, such simplifications are made by authors who do not bother to formalize the analysis and do not see the need to provide an exact list of included articles. This practice is more typical in more "liberal" or interdisciplinary fields, or where journals do not impose strict requirements.

Failure to comply with two or more requirements may also indicate a lack of awareness of common standards among authors and the absence of rigid review filters in relevant journals and conferences.

RQ2. Are there statistically significant differences in compliance between disciplines?

When analyzing the fulfillment of the requirements for SLRs in different disciplines, certain differences can be observed (Figure 3). For example, all four requirements are most often met in biochemistry (65%) and medicine (63%), and least often in computer science (38%). Conversely, in the first two fields it is extremely rare not to meet any of the requirements (1%), while in computer science it is not so rare anymore (10%). It should be noted that in this field, each requirement is fulfilled much less frequently than in the other fields.

This difference can be explained by the fact that the medical sciences have already established a "gold standard" – the PRISMA family of protocols – which prescribes these and other requirements for SLRs. Our study showed that in biochemistry and medicine,

80-85% of reviewed publications follow these protocols. It is so widespread that it has already penetrated deeply into many disciplines, including the social sciences (64%) and computer science (55%).

In the latter, an alternative methodology known as Kitchenham's guidelines (Kitchenham & Charters, 2007) is sometimes encountered (4%). Mentions of other methodologies occurred 1-2 times (totaling 1.5% of the sample) and were not

included in the analysis. Overall, this suggests that adherence to review across disciplines is related to the prevalence of the PRISMA standard.

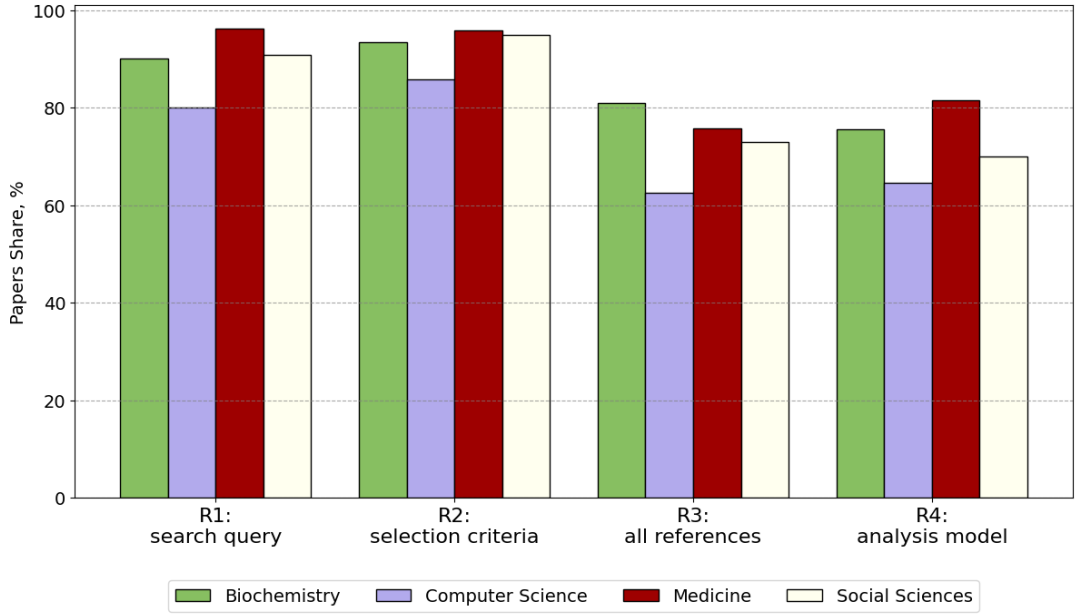

Figure 3. Degree of fulfillment of requirements by field of science

RQ3. Are these requirements more often fulfilled in high-impact journals?

At this stage of the study, high-impact journals are considered to be those that are in the top 10% of journals in a given scientific field according to the SJR (SCImago Journal Rank). Publications in other journals were used as a control group (Other). For each group, the proportion of publications that met all four requirements was calculated.

As can be seen in Figure 4, only in medicine were there significant differences: the high-impact journals met all requirements 13% more than the other journals (73% vs. 60%). This suggests that checking compliance with the requirements considered here (more precisely, the PRISMA requirements) is part of the editorial policy of leading medical journals.

In the other three areas, the difference is less than 2% – it is likely that the practice of strict adherence to systematic review methodologies has not yet taken hold in these areas, as journal editors do not prioritize it. In this case, an interesting phenomenon can be observed in biochemistry, where the PRISMA standard is recognized by the scientific community, but the editorial policies of leading journals are not affected. Another explanation could be that in these disciplines' other approaches (e.g. "mapping studies" or "narrative reviews") are used to prepare SRLs and therefore there is no need to insist on strict compliance with all formal criteria.

Thus, in medicine, high-impact journals play a more stringent "regulatory" role, ensuring that SLRs meet all criteria for methodological transparency.

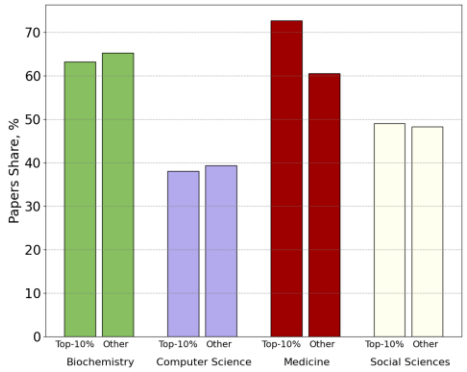

Figure 4. Percentage of SRLs that have all requirements met; comparing the 10% with the highest SJR to the others.

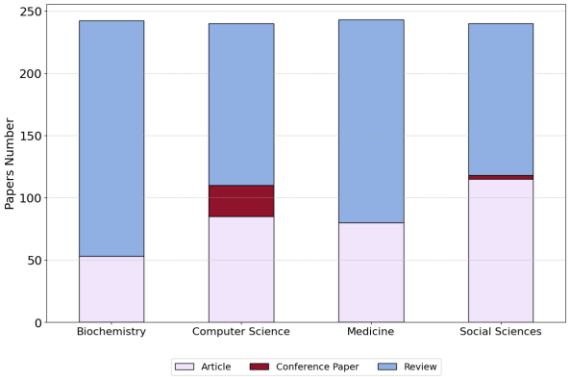

Figure 5. Distribution of SRLs by document type.

RQ4. Is there a relationship between completing requirements and citing SLRs?

After dividing the articles into groups according to the number of requirements fulfilled in them, we analyzed the distribution of field-weighted citation impacts obtained as of 2025/01/10. Contrary to our expectations, we found no significant differences in these groups, nor in those organized by discipline. It should be noted that the publications in question had a short "life cycle", whereas methodologically "high quality" papers may be recognized with a delay. However, we believe that methodological quality has little influence on the citation of the SLR, while more important factors are the relevance of the topic, the novelty or even the "brilliance" of the results, as well as the network of scientific communication and the authority of the authors. In the computer or social sciences, conceptual novelty, interdisciplinary scope, or practical implications may be more important than strict adherence to methodological guidelines. This is not to say that such reviews are not useful, but rather to distinguish between the notions of quality, relevance, and methodological rigor.

We feel it is necessary to highlight another important result. When searching for reviews in bibliographic review databases, a faceted filter by document type (doc_type=Review) is often used. Figure 5 shows that this results in filtering out 20 to 50% of publications that are also reviews, but of type Article. In addition, it is common practice in computer science to publish SLRs in conference proceedings with corresponding document types. There are also opposite situations where a document of type Review is not such a document. All this speaks not only about the imperfection of the mechanism of assigning document types in Scopus, but also about the mixing of two aspects in one doc_type attribute: source type (article for journals, CP for conferences, chapter for books) and content type (review, conference review, short survey, report). A complete solution to this problem is probably to separate these aspects into two different attributes and to clarify the rules for filling them in. Under the current conditions, we recommend not to filter by document type when systematically searching for reviews in Scopus.

Conclusion

Despite the widespread use of the PRISMA family of protocols, in practice there is still a certain "dis-synchronization" between what authors declare to be a "systematic review" and what is actually implied in the methodological guidelines. At the same time, the vast majority of authors already consider the search and inclusion criteria (R1, R2) as mandatory components of a SLR. However, a more detailed adherence to formal standards is not always realized, especially in fields with a less formalized methodological culture.

The presented results are preliminary. In the next phase of the study, we plan to expand the set of requirements under examination and to explore how their fulfilment relates both to the review methodologies employed and to the scope of the reference lists. The comprehensive list of requirements may eventually encompass all elements outlined in PRISMA – especially since *Frost (2022)* provides expert evaluation guidelines that could be adapted as prompts for LLMs. However, it should be noted that at present LLMs may not yet be able to thoroughly review all possible requirements, so the final set of criteria will need to be refined. A representative sample of SLRs will be peer reviewed using a similar methodology and the consistency of their results with the LLM data will be analyzed. As a result, the statistical significance of the results will be assessed. The project materials will be made available on GitHub.

References

Budgen, D., Brereton, P., Drummond, S. & Williams, N. (2018). Reporting systematic reviews: Some lessons from a tertiary study. *Information and Software Technology*, 95, 62–74.

Cooper, C., Booth, A., Varley-Campbell, J. *et al*. Defining the process to literature searching in systematic reviews: a literature review of guidance and supporting studies. *BMC Med Res Methodol* 18, 85 (2018). https://doi.org/10.1186/s12874-018-0545-3

Frost, A. D., Hróbjartsson, A., & Nejstgaard, C. H. (2022). Adherence to the PRISMA-P 2015 reporting guideline was inadequate in systematic review protocols. *Journal of Clinical Epidemiology*, 150, 179–187. https://doi.org/10.1016/j.jclinepi.2022.07.002

Hasan, B., Saadi, S., Rajjoub, N. *et al* (2024) Integrating large language models in systematic reviews: a framework and case study using ROBINS-I for risk of bias assessment *BMJ Evidence-Based Medicine* 2024;29:394-398. https://doi.org/10.1136/bmjebm-2023-112597

Kitchenham, B. & Charters, S. (2007) Guidelines for Performing Systematic Literature Reviews in Software Engineering, Technical Report EBSE 2007-001, Keele University and Durham University Joint Report.

Kitchenham, B., Madeyski, L., & Budgen, D. (2023). SEGRESS: Software Engineering Guidelines for REporting Secondary Studies. *IEEE Transactions on Software Engineering*, 49(3), 1273-1298.

Mathew, J.L. (2022) Systematic Reviews and Meta-Analysis: A Guide for Beginners. *Indian Pediatr* **59**, 320–330. https://doi.org/10.1007/s13312-022-2500-y

Norling B, Edgerton Z, Bakker C, Dahm P. (2021) The Quality of Literature Search Reporting in Systematic Reviews Published in the Urological Literature (1998-2021). *Journal of Urology*, 209(5), 837-843. https://doi.org/10.1097/JU.0000000000003190.

Tricco, A. C., Tetzlaff, J., Sampson, M. et al. (2008). Few systematic reviews exist documenting the extent of bias: A systematic review. *Journal of Clinical Epidemiology*, 61(5), 422–434.