Systematic Reviews

COMMENTARY

Open Access

Large language models, updates, and evaluation of automation tools for systematic reviews: a summary of significant discussions at the eighth meeting of the International Collaboration for the Automation of Systematic Reviews (ICASR)

Annette M. O'Connor[1*], Justin Clark[2], James Thomas[3], René Spijker[4,5], Wojciech Kusa[6], Vickie R. Walker[7] and Melissa Bond[3,8]

Abstract

The eighth meeting of the International Collaboration for the Automation of Systematic Reviews (ICASR) was held on September 7 and 8, 2023, at the University College London, London, England. ICASR is an interdisciplinary group whose goal is to maximize the use of technology for conducting rapid, accurate, and efficient evidence synthesis, e.g., systematic reviews, evidence maps, and scoping reviews of scientific evidence. In 2023, the major themes discussed were understanding the benefits and harms of automation tools that have become available in recent years, the advantages and disadvantages of large language models in evidence synthesis, and approaches to ensuring the validity of tools for the proposed task.

Keywords Automation tools, ChatGPT, Evidence synthesis, Large language models, Systematic reviews

*Correspondence:
Annette M. O'Connor
oconn445@msu.edu
[1] College of Veterinary Medicine, Michigan State University, East Lansing, MI 48824, USA
[2] Institute for Evidence-Based Healthcare, Bond University, Robina, QLD 4226, Australia
[3] EPPI Centre, UCL Social Research Institute, University College London, London WC1E 6BT, UK
[4] Julius Centre for Health Sciences and Primary Care, Cochrane Netherlands, University Medical Centre Utrecht, University Utrecht, Utrecht, the Netherlands
[5] Medical Library, Amsterdam Public Health, Amsterdam UMC, University of Amsterdam, Amsterdam, the Netherlands
[6] TU Wien, Vienna, Austria
[7] Division of Translational Toxicology, National Institute of Environmental Health Sciences, Research Triangle Park, Durham, NC, USA
[8] Knowledge Centre for Education, University of Stavanger, Stavanger, Norway

Introduction

This report summarizes the eighth meeting of the International Collaboration for the Automation of Systematic Reviews (ICASR), an interdisciplinary group with a shared interest in maximizing the use of technology to aid the transfer of scientific research findings to practice and to inform decision-making. ICASR aims to develop the capability for conducting accelerated, accurate, and efficient systematic reviews of scientific evidence. ICASR meetings have been held annually since 2017, except in 2020. Each meeting focuses on themes, and in 2023, the themes were as follows: (1) the benefits and harms of currently available automation tools, (2) the potential and pitfalls of large language models (LLMs) like ChatGPT for conducting systematic reviews, and (3) approaches to evaluation of automation tools that will increase trust and uptake.

Introduction and background

The International Collaboration for the Automation of Systematic Reviews (ICASR) was established in 2015 to enhance the efficiency and effectiveness of systematic reviews by integrating technologies like natural language processing and machine learning. The collaboration has hosted regular meetings to tackle diverse themes, share knowledge, boost collaboration, and shape the future of automated systematic reviews.

The inaugural meeting in Vienna in 2015 focused on establishing foundational principles for developing and integrating automation tools, known as the "Vienna principles." These principles emphasize the importance of efficiency, quality, continuous improvement, collaboration, and the open-source ethos in automating systematic reviews [1]. The second meeting in Philadelphia in 2016 expanded ICASR's scope, bringing together various stakeholders to identify challenges and propose short-term projects [2]. The focus was on enhancing understanding of available tools, developing validated datasets for tool testing, promoting tool interoperability, and establishing tool output quality criteria. London's 2017 meeting emphasized the immediate need for tool validation approaches and increased access to curated corpora. The participants outlined short-term goals, including publishing protocols for systematic review tasks and creating avenues for sharing corpora, for validation [3]. In 2018, the fourth gathering in The Hague explored the transferability of automation tools, the automated recognition of study designs, and strategies for evaluating these tools. This meeting was significant for recognizing the potential of tools developed for other purposes and their adaptability for systematic reviews [4]. The fifth meeting in Bergen in 2019 was themed around "information extraction from text," acknowledging the critical role of precise information extraction in enhancing the quality and speed of systematic reviews. Responding to the global pandemic, the sixth meeting in 2021 was conducted online, concentrating on "usability." Discussions delved into user experience, interoperability, tool evaluation, workflow integration, and the specific challenges and learnings presented by COVID-19 in the context of systematic reviews [5].

The seventh ICASR meeting, hosted in Köln by the German Institute for Quality and Efficiency in Health Care (IQWIG) in 2022, continued the tradition of annual gatherings. However, details of the meeting's discussions and outcomes were not specified in the provided text. The eighth and most recent meeting is the focus of this report. The University College London hosted the meeting on September 7 and 8, 2023. The focus of the conference was on progress in adopting automation, the role of LLMs in systematic reviews, the evaluation of automation, and the structure of ICASR for the next decade. While acknowledging the meaningful changes in systematic review workflow that have become available since 2015, the conference participants acknowledged the ongoing challenges, particularly for accurate automated data extraction and risk-of-bias assessment, evaluation, and the need for guidance from a group like ICASR. There was a consensus that while LLMs hold promise, their current utility in generating systematic reviews of scientific literature is limited.

Throughout these meetings, ICASR has promoted principles of automation of systematic reviews published as a consequence of the first meeting, i.e., the Vienna principles [6]. These principles advocate for a multifaceted approach to systematic review automation. They suggest a task-specific focus for automation, the potential for automation across all review stages, and the necessity for continuous process improvement. They stress maintaining high-quality standards of systematic reviews, promoting flexible tool use, encouraging collaboration among diverse groups, and committing to open-source code and accessible data sharing. These principles also underline the critical need for robust, replicable evaluation methods for automation tools, ensuring confidence and integrity in the systematic review process.

The efforts of ICASR to bring together multiple disciplines underscore a dedicated, evolving effort to leverage technology in systematic reviews, marking significant milestones while recognizing that the road ahead is filled with opportunities for groundbreaking advancements in this interdisciplinary field.

The eighth ICASR

Session 1: the benefits and harms of automation

Justin Clark, from Bond University, led the session as part of the first day of the meeting. This session emphasized the transformative impact of automation tools on systematic reviews. As predicted in 2015 at the formation of ICASR, many tools have become available for use by systematic reviews. Mr. Clark discussed several transformational tools created by his team and others. Justin Clark has transitioned in recent years from a user to a tool developer (though not a traditional software developer) and indicated, as many did, that successful tool development requires collaboration between reviewers and computer scientists. It was also pointed out that although many automation tools that are currently available are agnostic to the discipline of application, others have been developed for specific areas, usually health care, and still need translation to other disciplines such as social sciences, agricultural sciences, and environmental sciences.

Justin Clark discussed the development process, including the initial task breakdown for systematic reviews. This meticulous breakdown helped identify tasks suitable for automation, particularly those considered tedious or less preferred by the team. One limitation noted was that computers still struggle with data extraction despite some advancements.

Specific tools were showcased, including the following:

1. The Polyglot Search Translator [7]: This tool translates search queries across various databases, significantly reducing the time spent on manual translations. The unexpected advantage was the extra time to refine search strategies, resulting in fewer, more targeted results and less screening effort. Approximately, 50% less screening was needed due to more precise searches. Polyglot is an example of a tool generally designed for bibliographic databases relevant to human clinical and public health. Translating other bibliographic databases is a potential future project with the appropriate collaborators.
2. RevMan Replicant [8]: This tool streamlines the writing of the results section of reviews by extracting data from RevMan 5 files. This automated writing tool has helped eliminate the daunting task of writing into a blank document and significantly reduced transcription errors by direct content generation, enhancing efficiency and accuracy. This tool is available for any review team that uses RevMan. However, familiarity with RevMan as a systematic review tool is often restricted to research synthesizers working in human clinical and public health.

The key take away is that these automation tools while enhancing efficiency have been effective and are readily available. However, the tools do not replace the fundamental skills required to conduct systematic reviews. The tools aid in task execution, enabling more focus on critical review aspects, thereby improving overall review quality. Moreover, it was stressed that these tools are free to access, encouraging widespread use and skill development in systematic reviews.

Session 2: using large language models in systematic reviews — potential vs. pitfalls

A specific session was devoted to large language models (LLMs) and their use in systematic reviews. In 2022, ChatGPT was released publicly and transformed the access of review teams to these models, which had previously been accessible only to a select group of researchers. The no-code interface has resonated with users, and enormous possibilities are envisioned for these models. However, utilizing LLMs in the systematic review workflow is a double-edged sword, presenting remarkable opportunities and significant challenges. This session aimed to discuss these models openly, understand the views, and raise questions about their current usage, especially in systematic reviews requiring utmost accuracy and factual integrity. Three speakers discussed LLMs.

LLMs in systematic reviews: Potential vs. Pitfalls Dr. I. Marshall, from Kings College London, pointed out that while LLMs have advanced, their application within systematic reviews is not without risk. The ability of LLMs to generate coherent and fluent text often overshadows the factual inaccuracies included in the text. The crux lies in their training on vast, unlabeled datasets, making validating their outputs complex and challenging. This is particularly precarious in multi-document summaries, a staple in systematic reviews, where consistency and accuracy are paramount. LLMs struggle with this aspect in their current form, sometimes yielding synthesized content that lacks factual correctness.

The dynamic nature of these models further complicates their utility for reproducible systematic reviews. The constant updates and iterations these models undergo make reproducing results nearly impossible at present, presenting a significant hurdle in environments that rely on consistency and traceability. Additionally, there is a troubling aspect of accountability; pinpointing responsibility for misinformation remains nebulous due to the models' opaque functioning. Despite these concerns, LLMs hold promise for tasks like drafting initial summaries or generating templates, potentially

expediting the systematic review process. However, their limitations necessitate a cautious approach backed by robust evaluation mechanisms.

Enhancing LLMs with cognitive factoring and collaborative techniques One suggested solution to mitigate LLM's limitations is "factored cognition." Representatives of the company, Elicit (https://elicit.com) introduced this concept in their presentation, emphasizing breaking down complex questions into more straightforward, manageable tasks. This method aims for a collaborative problem-solving approach between AI and humans, enhancing LLMs' reliability.

Factored cognition could address concerns about how LLMs handle systematic review tasks by dissecting broad questions into specific prompts, allowing for more accurate, manageable responses. The presenter proposed that factored cognition solved the "hallucination" issue, where LLMs often produce inaccurate or fabricated data while attempting to generate informative responses. However, evidence to support this statement was not provided.

By integrating LLMs with cognitive strategies, we can potentially harness efficiency while minimizing the propensity for error. However, this area is still under exploration, and its practical efficacy remains to be thoroughly evaluated.

Refining systematic review development of citation searches with advanced prompting systems Mr. H. Scells reported research on using LLMs for systematic review queries. After a brief discussion of the conceptual [9] and objective [10] approaches to identifying search terms, Mr. Scells presented the results of a study using ChatGPT for such searches [11]. The study used two previously published review study test collections: CLEF TAR [12] and their seed study test collection [11]. The study used unguided and guided prompt approaches to developing the searches. The findings highlighted that the quality of results is heavily contingent on the prompts, with guided prompts yielding more reliable outputs than broad, unguided prompts [11]. An important finding was the LLMs' tendency to misconstrue certain terms, which affected search precision. To counter this, the suggestion was to integrate external databases like MeSH with ChatGPT as a leap towards informed AI use, ensuring the model's outputs are anchored in verifiable data. Finally, as mentioned previously, the variation in responses to ChatGPT prompts is a problematic feature that limits the reliability of these methods currently for search design. The conclusion was that LLMs in their current state require supervision by experts versed in the task (e.g., information specialists) to ensure their generated queries align with the stringent demands of scientific research. The potential for misuse or misinterpretation is high if these systems are the starting point for the search query design process if used without proper guidance or foundational data.

The conclusion from "Session 2: Using large language models in systematic reviews — potential vs. pitfalls," focusing on LLMs, is that LLMs offer transformative potential for systematic reviews, promising efficiency and innovation. However, their integration necessitates a paradigm shift, acknowledging their limitations, ethical implications, and the critical need for human oversight. Future directions are inclined towards promoting evaluations and applications with open-source models, enabling more transparent, reproducible research. But until then, a balanced viewpoint is essential, advocating for the cautious and informed application of LLMs, ensuring they serve as reliable aids in the complex landscape of systematic reviews rather than unpredictable variables.

Session 3: evaluation of automation tools — what has been done and what is needed?

A comprehensive dialogue on evaluating automation tools for systematic reviews dissected numerous facets of the topic, presenting research findings, contemplating methodologies, and forecasting future necessities and trends. The session was anchored around the evaluation of automation tools, particularly their evaluation, effectiveness, and enhancement in various sectors, with a spotlight on health and climate change.

The discussion commenced with an in-depth look at a current Wellcome Trust-funded project examining the use and reporting of automation tools within climate change and health [13]. This multiphase initiative seeks to identify and critically appraise evaluations of automation tools for evidence synthesis (called digital evidence synthesis tools, or DESTs, within the context of the project) and consider their applicability to the field of climate and health. The project team has published a "map" of automation tool evaluations (available at the previous link). The team will also publish a mapping review exploring the use and reporting of DESTs within climate and health evidence syntheses. The project will also include a series of case studies, where current practices and tools will be examined in depth: large-scale tools for automatically mapping the research literature, an exploration of how LLMs have been used to assist with screening and data extraction, and their potential application in climate and health, as well as the potential of tools using LLMs to assist in evidence synthesis at scale. The project has identified a critical inconsistency in reporting standards

across disciplines, which hampers automation efforts for diverse topics like climate change. This discrepancy poses a significant challenge to the effective utilization of automation tools, compounded by apprehensions about the requisite for specialized knowledge in AI and machine learning.

The project's forthcoming phase aims to develop recommendations for refining these digital tools' relevance and applicability in diverse contexts. This approach will use extensive feedback from detailed case studies and stakeholders, including researchers, tool developers, and policymakers, contributing pragmatic insights into these tools' performance in different settings. Significant emphasis was on transitioning traditional publication systems to more computable, semantic ecosystems, underscoring the difficulty in handling non-machine-readable documents and recognizing gaps in capturing emerging technologies like LLMs.

Delving into the evaluation specifics, one presentation by Dr. M. Bond illuminated an unsettling revelation: a very small percentage of systematic reviews openly disclose the use of advanced AI technologies, with a surprising reliance on rudimentary tools, both within the fields of education [14] and climate and health [15]. The project's methodology whittled down hundreds of potential evaluations to a concentrated selection, focusing on genuine, in-action assessments of tools within systematic reviews. Though centered predominantly around specific tools, these evaluations unearthed additional tools or algorithms, indicating a richness in the landscape not fully explored.

Trust in these digital tools emerged as a paramount theme. Anecdotal evidence suggested a noticeable hesitation towards complete automation (e.g., [16]), particularly from transient initiatives like academic research projects.

The dialogue wove through various considerations, from the potential of domain-specific language models, the balance between early-stage research and mature AI products, to the responsibility dichotomy between tool developers and users. It touched upon the essential need for education on AI tool usage, foreseeing integration at undergraduate levels to cultivate informed application in evidence synthesis [17].

A recurrent concern was the sustainability of AI tools post-funding, spotlighting the critical role of consistent financial and policy support in maintaining tool efficacy. The conversation culminated in a reflection on the cyclical dilemma of tool adoption and funding: tools must demonstrate the effectiveness of financing and sustainability but simultaneously require financial injection for comprehensive evaluation and improvement.

The discussion also broached the urgency within academic circles to adopt AI tools, even prematurely or inappropriately, driven by availability rather than suitability. This tendency could cement reliance on potentially subpar tools and a potentially dangerous normalization of inaccuracy, underscoring the requirement for stringent quality control in systematic reviews.

In a detailed sidebar, the participants debated the intricacies of evaluating systematic review automation, focusing on traditional evaluation metrics and the impact of review outcomes. They discussed the limitations of the Work Saved over Sampling (WSS) metric and advocated for the use of the true-negative rate (TNR) for better comparability [18]. They endorsed more intuitive measures that could offer consistent, standardized analyses across diverse datasets and models and experimented with custom evaluation metrics, shedding light on their performance nuances. An outcome-based evaluation framework has also been proposed [19]. This approach moves beyond mere recall rates to consider the actual influence of studies on the final outcomes of systematic reviews, arguing that not all studies contribute equally to a review's conclusions.

The last presentation addressed the need for statistically validated stopping criteria in machine learning for document screening, emphasizing the use of probability theory to determine when to cease the screening [20]. It highlighted the importance of incorporating adjustments for bias and the necessity of integrating these criteria into machine learning platforms to enhance the efficiency and reliability of systematic reviews.

In the grander scheme, the conversation encapsulated a profound exploration of the future of systematic reviews in AI and machine learning. The participants discussed the challenges current systematic review datasets curators face, including documentation, availability, and size issues. In response, the CSMeD benchmark has been proposed to standardize evaluations, provide unified API access, and bridge gaps in current datasets [21]. The discourse underscored the need for methodological enhancements, sophisticated analyses, standardized evaluation metrics, and a harmonious balance between technological advancement and human insight. This equilibrium would ultimately shepherd the field towards a future where digital evidence synthesis tools are advanced and intuitively aligned with human reasoning and industry requirements.

Session 4: the future of ICASR

Since the first meeting in 2015, members of ICASR have seen substantial growth in the number and use of automation tools, e.g., Covidence reports over 300,000 users have used it to start over 265,000 reviews (https://www.

covidence.org/). This growth in the use of automation tools raises the inevitable issue of whether the tools are safe to use or if the speed increase causes a quality reduction. This has been particularly noticeable to the longer-term members of ICASR, who meet and collaborate informally.

This has highlighted the lack of a guiding body that can help ensure standardized, best-practice evaluation methods are used to provide the robustness of the automation tools being released. Therefore, one crucial note to come from the meeting was the need to set up this sort of group.

To begin the process of a more structured organization with resources to support the needs of the automated evidence synthesis community, ICASR will seek funding from bodies interested and supportive of this area. Initially, this funding will be directed towards two goals: (1) improving how information is communicated around automation research and (2) identifying and collating data that can be used in evaluation research. These were seen as the most significant and critical initial issues to overcome. With more robust communication and evaluation data in place, ICASR will design and write guidance documents to evaluate the multiple tasks required to synthesize existing evidence accurately and transparently. It is planned that the first steps will be in place before the next meeting in late 2024, with the design of evaluating guidance as a focus of ICASR 2024.

Conclusion

With the advances in research and tools for the synthesis of evidence, it has become clear that guidance on evaluating these advances is sorely needed. As guidance is lacking and no other body is in a position to offer this guidance, ICASR has decided they need to take up this burden.

Acknowledgements
The authors thank the ICASR 2023 participants for their active and thoughtful discussions.

Authors' contributions
AOC prepared the first draft of the manuscript. All authors (AOC, JC, JT, RS, WK, VRW, and MB) participated in its review and revision. Publications costs were met by AOC. All authors read and approved the final manuscript.

Funding
Not applicable.

Data availability
Not applicable.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 16 July 2024 Accepted: 24 September 2024
Published online: 27 November 2024

References

1. Beller E, Clark J, Tsafnat G, Adams C, Diehl H, Lund H, et al. Making progress with the automation of systematic reviews: principles of the International Collaboration for the Automation of Systematic Reviews (ICASR). Syst Rev. 2028;7(1):77. https://doi.org/10.1186/s13643-018-0740-7.
2. O'Connor AM, Tsafnat G, Gilbert SB, Thayer KA, Wolfe MS. Moving toward the automation of the systematic review process: a summary of discussions at the second meeting of International Collaboration for the Automation of Systematic Reviews (ICASR). Syst Rev. 2018;7(1):3. https://doi.org/10.1186/s13643-017-0667-4.
3. O'Connor AM, Tsafnat G, Gilbert SB, Thayer KA, Shemilt I, Thomas J, et al. Still moving toward automation of the systematic review process: a summary of discussions at the third meeting of the International Collaboration for Automation of Systematic Reviews (ICASR). Syst Rev. 2019;8(1):57. https://doi.org/10.1186/s13643-019-0975-y.
4. O'Connor AM, Glasziou P, Taylor M, Thomas J, Spijker I, Wolfe MS. A focus on cross-purpose tools, automated recognition of study design in multiple disciplines, and evaluation of automation tools: a summary of significant discussions at the fourth meeting of the International Collaboration for Automation of Systematic Reviews (ICASR). Syst Rev. 2020;9(1):100. https://doi.org/10.1186/s13643-020-01351-4.
5. Available from: http://ojs.eahil.eu/ojs/index.php/JEAHIL/issue/view/187. Cited 2023 Nov 20.
6. Beller E, Clark J, Tsafnat G, Adams C, Diehl H, Lund H, et al. Making progress with the automation of systematic reviews: principles of the International Collaboration for the Automation of Systematic Reviews (ICASR). Syst Rev. 2018;7(1):77. https://doi.org/10.1186/s13643-018-0740-7.
7. Kung J. Polyglot search translator. J Can Health Libr Assoc. 2022;43(1):35–9. https://doi.org/10.29173/jchla29600.
8. Institute for Evidence-Based Healthcare. RevMan Replicant. Available from: https://github.com/IEBH/revman-replicant. Cited 2023 Nov 23.
9. Clark J. Systematic Reviewing. In: Doi S, Williams G. (editors), Methods of clinical epidemiology. Springer Series on Epidemiology and Public Health. Berlin, Heidelberg: Springer; 2013. https://doi.org/10.1007/978-3-642-37131-8_12.
10. Hausner E, Waffenschmidt S, Kaiser T, Simon M. Routine development of objectively derived search strategies. Syst Rev. 2012;1: 19. https://doi.org/10.1186/2046-4053-1-19.
11. Wang S, Scells H, Koopman B, Zuccon G. Can ChatGPT write a good Boolean query for systematic review literature search? In: Proceedings of the 46th international ACM SIGIR conference on research and development in information retrieval (SIGIR '23). New York; 2023. pp. 1426–1436. https://doi.org/10.1145/3539618.3591703.
12. Kanoulas E, Li D, Azzopardi L, Rene Spijker R. CLEF 2018 technologically assisted reviews in empirical medicine overview. In: CEUR Workshop Proceedings. 2125. Invited Paper 6. Available at: https://ceur-ws.org/Vol-2125/invited_paper_6.pdf.
13. Bond M, Finnerty A, O'Mara-Eaves A, O'Driscoll P, Thomas J, Minx J, et al. Digital evidence synthesis tool evaluations. 2024. EPPI Visualiser database. Available at: https://eppi.ioe.ac.uk/eppi-vis/login/open?webdbid=435. Cited 2024 May 22.
14. Bond M, Khosravi H, Bergdahl N, Buntins K, De Laat M, Oxley E, et al. Digital evidence synthesis tools in educational technology research: a systematic mapping review. 2023. Pre-print. https://doi.org/10.13140/RG.2.2.30594.25288.
15. Bond M, O'Mara-Eaves A, O'Driscoll P, Thomas J, Minx J, Callaghan M, et al. 2024. Digital evidence synthesis tool use in climate and health. EPPI Visualiser database. Available at: https://eppi.ioe.ac.uk/eppi-vis/login/open?webdbid=505. Cited 2024 May 22.

16. Jardim PSJ, Rose CJ, Ames HM, Echavez JFM, van de Velde S, Muller AE. Automating risk of bias assessment in systematic reviews: a real-time mixed methods comparison of human researchers to a machine learning system. BMC Med Res Methodol. 2022;22(1):167. https://doi.org/10.1186/s12874-022-01649-y.
17. Bond M, Khosravi H, de Laat M, Bergdahl N, Negrea V, Oxley E, Pham P, Chong SW, Siemens G. A meta systematic review of artificial intelligence in higher education: a call for increased ethics, collaboration, and rigour. Int J Educ Technol High Educ. 2024;21(1):Article Number 4. https://doi.org/10.1186/s41239-023-00436-z.
18. Kusa W, Lipani A, Knoth P, Hanbury A. An analysis of work saved over sampling in the evaluation of automated citation screening in systematic literature reviews. ISWA. 2023;2023(18):200193. https://doi.org/10.1016/j.iswa.2023.200193.
19. Kusa W, Zuccon G, Knoth P, Hanbury A. Outcome-based evaluation of systematic review automation. In: Proceedings of the 2023 ACM SIGIR international conference on theory of information retrieval (ICTIR '23). New York, 2023. p. 125–133. https://doi.org/10.1145/3578337.3605135.
20. Callaghan MW, Müller-Hansen F. Statistical stopping criteria for automated screening in systematic reviews. Syst Rev. 2020;9(1):273. https://doi.org/10.1186/s13643-020-01521-4.
21. Kusa W, Mendoza OE, Samwald M, Knoth P, Hanbury A. CSMeD: bridging the dataset gap in automated citation screening for systematic literature reviews. In: Thirty-seventh conference on neural information processing systems datasets and benchmarks track. 2023. p. 17. Available at: https://openreview.net/pdf?id=ZbmS3MU25p.

Publisher's Note