


RESEARCH

Open Access



# Delphi survey on the most promising areas and methods to improve systematic reviews' production and updating

Mersiha Mahmić-Kaknjo<sup>1,2\*</sup> , Vicko Tomić<sup>3,4</sup>, Moriah E. Ellen<sup>5,6</sup>, Barbara Nussbaumer-Streit<sup>7</sup>, Raluca Sfetcu<sup>8,9</sup>, Eduard Baladia<sup>10</sup>, Nicoletta Riva<sup>11</sup>, Angelos P. Kassianos<sup>12,13</sup> and Ana Marušić<sup>4</sup>

## Abstract

**Background** Systematic reviews (SRs) are invaluable evidence syntheses, widely used in biomedicine and other scientific areas. Tremendous resources are being spent on the production and updating of SRs. There is a continuous need to automatize the process and use the workforce and resources to make it faster and more efficient.

**Methods** Information gathered by previous EVBRES research was used to construct a questionnaire for round 1 which was partly quantitative, partly qualitative. Fifty five experienced SR authors were invited to participate in a Delphi study (DS) designed to identify the most promising areas and methods to improve the efficient production and updating of SRs. Topic questions focused on which areas of SRs are most time/effort/resource intensive and should be prioritized in further research. Data were analysed using NVivo 12 plus, Microsoft Excel 2013 and SPSS. Thematic analysis findings were used on the topics on which agreement was not reached in round 1 in order to prepare the questionnaire for round 2.

**Results** Sixty percent (33/55) of the invited participants completed round 1; 44% (24/55) completed round 2. Participants reported average of 13.3 years of experience in conducting SRs (SD 6.8). More than two thirds of the respondents agreed/strongly agreed the following topics should be prioritized: extracting data, literature searching, screening abstracts, obtaining and screening full texts, updating SRs, finding previous SRs, translating non-English studies, synthesizing data, project management, writing the protocol, constructing the search strategy and critically appraising. Participants have not considered following areas as priority: snowballing, GRADE-ing, writing SR, deduplication, formulating SR question, performing meta-analysis.

**Conclusions** Data extraction was prioritized by the majority of participants as an area that needs more research/methods development. Quality of available language translating tools has dramatically increased over the years (Google translate, DeepL). The promising new tool for snowballing emerged (Citation Chaser). Automation cannot substitute human judgement where complex decisions are needed (GRADE-ing).

**Trial registration** Study protocol was registered at <https://osf.io/bp2hu/>.

**Keywords** Evidence synthesis, Automation tools, Prioritization

\*Correspondence:

Mersiha Mahmić-Kaknjo  
mmahmickaknjo@gmail.com

Full list of author information is available at the end of the article



© The Author(s) 2023. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

## Background

Systematic reviews (SRs) are evidence syntheses that serve as valuable support for decision-making in health-care [1, 2] and social sciences [3, 4]. They can be defined as a summary of studies addressing a specific topic using reproducible analytical methods to collect secondary data and analyse it using systematic and explicit methods to identify, select and critically appraise relevant studies, and to extract and summarize the data [5]. SRs can be used to reduce biases and provide evidence to stakeholders such as policymakers, decision-makers, practitioners, researchers, academia, the public, and citizens [6].

Since the 1990s, when organizations like Cochrane, Campbell Collaboration, and the Joanna Briggs Institute (JBI) emerged [1, 7, 8], there has been an increase in both the number of SRs and their utilization to inform policy and practice [9]. Due to the fact that tremendous resources are needed to produce and update SRs, there is a need to automatize the process as much as possible and use the workforce and resources to make it more efficient [10, 11]. This study was conceptualized within the framework of Evidence-Based REsearch (EVBRES) [12], which is a 4-year (2018–2022) EU-funded COST Action CA-17117 with over 40 countries participating globally, aiming to encourage researchers and other stakeholders to use an Evidence-Based Research (EBR) approach while carrying out and supporting clinical research—thus avoiding redundant research. The Action has been extended until 16th April 2023, and as part of the research agenda of a working group (WG3) focusing on improving efficiency in producing and updating systematic reviews [10, 13, 14]. Based on the results of previous activities, we designed a Delphi study (DS) to reach an agreement on prioritising the most promising areas and methods to improve the efficiency of producing and updating SRs. DSs offer a flexible approach to obtaining information regarding how best to allocate professional resources such as knowledge and expertise [15, 16]. This is especially important when the agreement on statement is needed based on the best available evidence. In this study, the expert consensus regarding the most promising areas and methods to improve the efficient production and updating of SRs was pursued.

## Methods

A DS was employed to gain expert insight regarding which areas and methods need to be prioritized to improve efficiency in producing and updating SRs. The DS was conducted exclusively online since face-to-face interaction was neither preferred [15] nor achievable due to Coronavirus disease (COVID-19) pandemic travel restrictions. Recommendations for Conducting and

Reporting Delphi Studies (CREDES) [17] were followed throughout the manuscript, excluding parts that were beyond the scope of this project (i.e., external validation and dissemination). Participants were provided with a description of the overall aim of the EVBRES and the specific objective of the DS.

Usually, data from the first round of a DS are solely qualitative [18], but as this survey was informed by a scoping review [10] and a qualitative study [19], the quantitative techniques could be applied as early as in round 1.

Round 1 of the DS was launched on November 15th, 2021, and two reminders were sent 2 weeks apart from each other. Round 1 ended on December 14th, 2021. Round 2 of the DS was launched on April 11th, 2022, and three reminders were sent two weeks apart from each other. The survey ended on May 25th, 2022. Since our study was thoroughly informed by previous EVBRES research [10, 11, 13], agreement was expected to be reached after two rounds. As agreement was reached after round 2, there was no need for round 3, and we closed the DS.

## Participants

As response characteristics of a small expert panel in a well-defined knowledge area are considered reliable in light of augmented sampling [20], 55 experienced authors of SRs were invited using a combination of two non-probabilistic sampling methods, both widely used in Delphi methods [21]: purposive sampling technique followed by criterion sampling in which members of the EVBRES were requested to provide us with personal contacts that met the inclusion criteria for participation. Potential participants were contacted by an email from the EVBRES member and informed what they would be asked to do, how much time they would be expected to contribute, and when, as well as what use would be made of the information they provided. Inclusion criteria were (a) participating in at least 3 SRs as an author; and (b) being a first/last/senior/mentor author in at least one SR. All suggested participants' names were searched using Google Scholar [22] prior of sending invitations in order to confirm that they satisfied the inclusion criteria.

## Delphi survey design and procedures

### Survey design

To design the surveys, seven formulation and review sessions were conducted with at least two research team members (1 permanent and 2 alternating research team members). LimeSurvey Professional (LimeSurvey Cloud 3.27.24, LimeSurvey GmbH, Hamburg, Germany) was used to design the online survey.

**Table 1** Organization scheme of DS questionnaire

Section	Type of data	Variables
Demographic data	Range, not an exact value; in order to minimize a chance to identify participants	Age, gender, years of experience in conducting SRs, number of conducted SRs, number of SRs that they have led, role/s in conducting SRs, and area of employment
Prioritisation	(1) 5-point Likert scale mandatory question regarding whether the step is time/effort/resource-intensive and should be prioritized in future research concerning methods of development and automation Participants were offered fixed statements that a particular step needs to be prioritized in future research and an open-ended field. For each statement, participants had to rate how strongly they agree that the topic is important to include (1—“strongly disagree”, 2—“disagree”, 3—“indifferent”, 4—“agree” and 5—“strongly agree”, “I do not know”) (2) Participants were encouraged to provide arguments for the ratings through open responses	Steps of SR production: (1) project management, (2) formulating the review question, (3) finding previous SRs, (4) writing the protocol, (5) constructing the search strategy, (6) literature searching, (7) de-duplicating, (8) screening abstracts, (9) obtaining full-text, (10) screening full-texts, (11) snowballing-citation chasing/tracking, (12) translating non-English studies into English, (13) extracting data, (14) critically appraising, (15) synthesizing data, (16) GRADE-ing ( <a href="https://www.gradeworkinggroup.org/">https://www.gradeworkinggroup.org/</a> )—going from evidence to decision, (17) updating the review to see whether some new studies were published between the search date and the final version of the article, (18) performing a meta-analysis, and (19) writing up the review*
Qualitative section	Open-ended long text field (non-mandatory question) in which participants were invited to freely discuss any issue they find important in that step. Participants were encouraged to provide as many opinions as they felt appropriate	(1) How the methodology of producing SRs can be improved, (2) areas that should be prioritized in future research, and (3) other issues considered important regarding SRs’ production and updating

\* Based on work by Tsafnat et al. [23] as well as our team’s previous work [11]

Round 1 of the DS included sections presented in Table 1.

The study was piloted among 10 participants of the EVBRES on June 5th, 2021 and some minor adjustments were made to make the questionnaire more user-friendly and time efficient.

In round 2, areas where agreement was not reached in round 1 were further explored by testing statements with required 5-point Likert answers. This included 2 statements on “snowballing” (“development of better tools is needed” and “automation can be helpful in this area”), 4 statements on “GRADE-ing” (“this step is methodologically well developed”, “potential for automation is low in this step”, “standardization of GRADE assessment may be helpful”, “this step is relatively low resource task”) and 2 statements on “deduplication” (“automation is advanced in this step”, “there is scope for improvement”). Theme “meta-analyzing” was not further explored since participants commented that they considered that it refers to “synthesising data”.

#### Data analysis

For the prioritisation exercise 5-point Likert-scale based quantitative answers was analysed. When analysing the questions regarding prioritization, it was considered that consensus was reached when two-thirds or more (66.7%) of the participants’ responses reached a certain score range (“agree” and “strongly agree”, or “disagree” and “strongly disagree”). Since our sample was small and skewed interquartile range (IQR) as a measure of variability and median as a measure of central tendency were

chosen in order to find out in which range the most of the results lie: smaller the values, less skewed the results.

A reflexive thematic analysis approach was used to analyse the qualitative data [24, 25] since the theoretical freedom of this approach allows flexibility. Following the familiarization with the data through transcription, reading, and re-reading, initial codes were generated and gathered into potential themes. After reviewing themes across an entire data set, a thematic map was developed, and clear definitions and names for each theme were identified and further refined through ongoing analysis. The data were coded by one author (VT) with an inductive approach and themes were developed at a semantic level. Concepts of data or code saturation were not used in this study because they were not consistent with reflexive thematic analysis values and assumptions [24]. Taxonomy was developed manually and independently at the same time by another researcher (MMK), and the taxonomy choice finalized by the third researcher (AM). Minimal grammar and spelling corrections were made to participants’ answers by a researcher that is a native English speaker (MEE). In round 2 of the DS, a taxonomy based on the responses from the first surveys was presented to participants.

Data gathered with the online survey system were exported to Microsoft Excel and SPSS 16+ compatible files and then analysed with NVivo 12 Plus for Windows (QSR International Pty Ltd., London, UK), Microsoft Excel 2013 (Microsoft Corporation, Redmond, WA, USA) and IBM SPSS Statistics for Windows, version 28.0 (IBM Corp., Armonk, NY, USA).

## Results

A total of 55 participants were invited and 39 agreed to participate. In round 1, 33 completed the survey in its entirety, 3 participants opted out, and 3 participants did not provide complete responses. These 33 participants were invited to the round 2, and 24 of them completed it. Since the consensus was reached after sending 3 reminders 2 weeks apart from each other, the survey was ended.

### Participants' characteristics

Of the 33 participants who completed the questionnaire in round 1; 14 identified as male, 17 as female, and 2 of them preferred not to specify their gender. There were representatives of all age groups, but most participants were aged 41–50 years ( $n=13$ ), or 31–40 years old ( $n=9$ ). Since the average experience in conducting SRs was 13.33 years, it can be stated that our participants were experts in the field (Fig. 1).

### Descriptive data analysis

For the general Likert-scale questions in round 1, with regards to identifying which areas and methods of the process of conducting SR are most time/effort/resource-intensive and should be prioritized in future research concerning methods of development and automation, consensus was reached (more than 66.7% participants agreed or strongly agreed that topic should be prioritized) for 13 out of the 19 topics, as is shown in Fig. 2. Data extraction was prioritized by the majority as an area where research and automation can help reduce the intensity of resource use; 90% (30 out of 33) have “strongly agreed” or “agreed”; which was also emphasized in qualitative results section. Screening abstracts has the most “strongly agree” answers 51% (17 out of 33) and by adding “agree” it comes third 82% (27 out of 33). Great majority of participants (27 out of 33; 82%) prioritized literature searching (12 participants “strongly agree” and 15 “agree”), obtaining full texts (11 participants “strongly agree” and 16 “agree”) and updating the SR (10 “strongly agree” and 17 “agree”).

As depicted in Fig. 3, during round 2, participants reached a consensus on all 8 tested statements with 66.7% either agreeing/strongly agreeing or disagreeing/strongly disagreeing. The most supported statements from round 2 are for snowballing and GRADE-ing: 83% or 20 out of 24 participants reached consensus (10 “strongly agree”, 10 “agree”) that automation can be helpful in snowballing; 79% or 19 out of 24 reached consensus (9 “strongly agree”, 10 “agree”) that GRADE-ing is a complex task that requires human judgement and potential for automation is low in this activity. The latter consensus of 79% was

also reached regarding statements that GRADE-ing is methodologically well developed (9 “strongly agree” and 9 “agree”), as well as development of better tools in snowballing (6 “strongly agree” and 13 “agree”).

### Qualitative data analysis

The following three main themes were developed from the qualitative data gathered in the DS: (1) the most important tools and approaches, (2) different areas and methods require different levels of automation, and (3) prioritization concerning future research of particular methods is crucial to improve efficiency (Fig. 4).

The first theme (Table 2) developed in our qualitative study is regarding what are considered the most important tools and approaches used to produce and update SRs. Participants mostly pointed out tools and approaches that are already available on the market but are often not known or used by most researchers. However, they also recommended some tools and approaches that need to be developed to improve the efficiency of SRs. As a result, two sub-themes were developed: “Existing tools and approaches that can improve efficiency” and “Tools and approaches that need to be developed”.

For areas such as “Formulating the review question” and “Writing the protocol”, the participants recommended SUAMRI, IEBHC Review Wizard, and RevMan. They also pointed out several other tools developed by the Institute for Evidence-Based Healthcare at Bond University in Australia: Systematic Review Accelerator (SRA), PRISMA for Abstracts, TIDieR, Shared decision making, EBM teaching resources, CriSTAL Tool, and MASCoT.

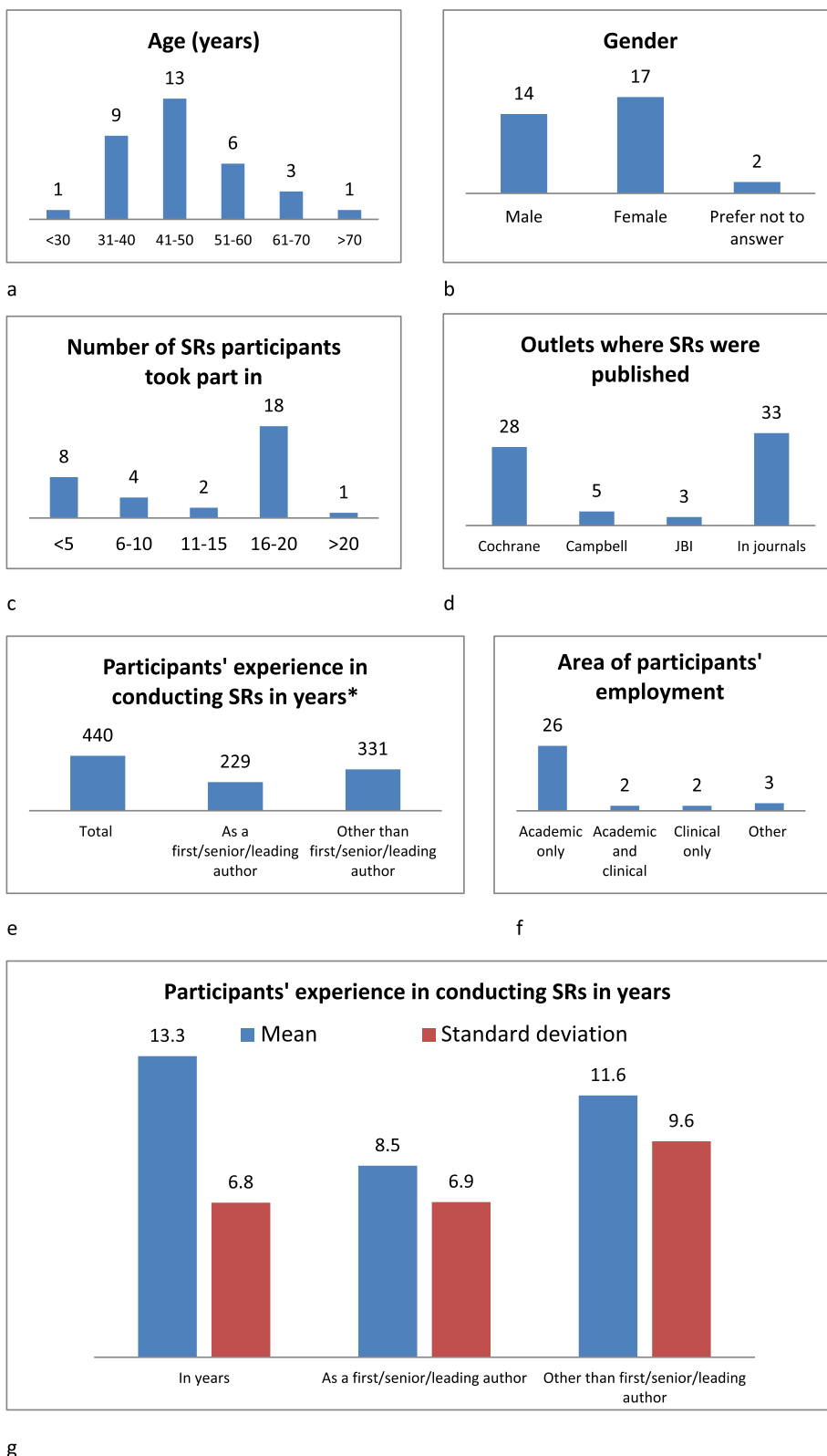
Regarding de-duplicating, screening abstracts, and full-texts, the expert suggested that tools such as EPPI-Reviewer, Covidence, DistillerSR can improve the efficiency of SRs.

Some participants pointed out that some handy tools already exist, such as “Citation chaser” (<https://estech.shinyapps.io/citationchaser/>) that can be used for citation tracking.

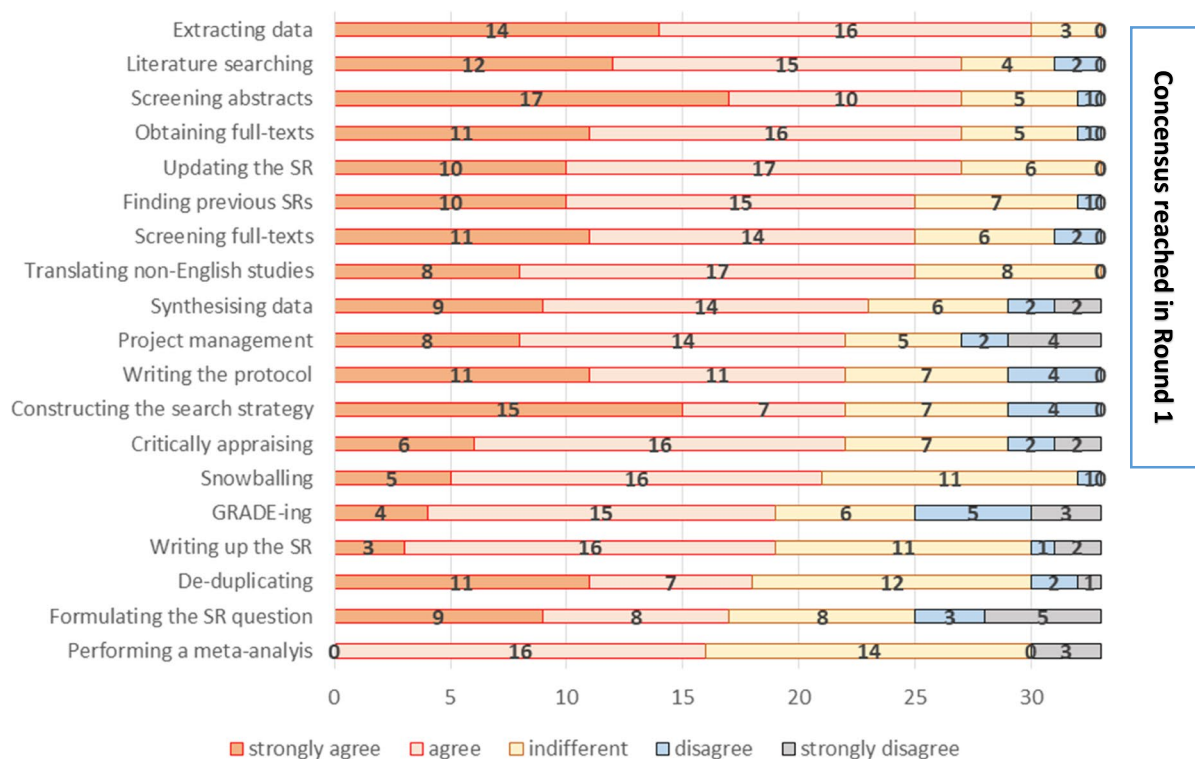
The participants also pointed out that the quality of available tools for translating non-English studies has increased dramatically over the years, suggesting that Google Translate and DeepL seem to be the most useful tools in this area.

Additional training on how to use tools and training in general was seen as something that could improve the overall quality of SRs.

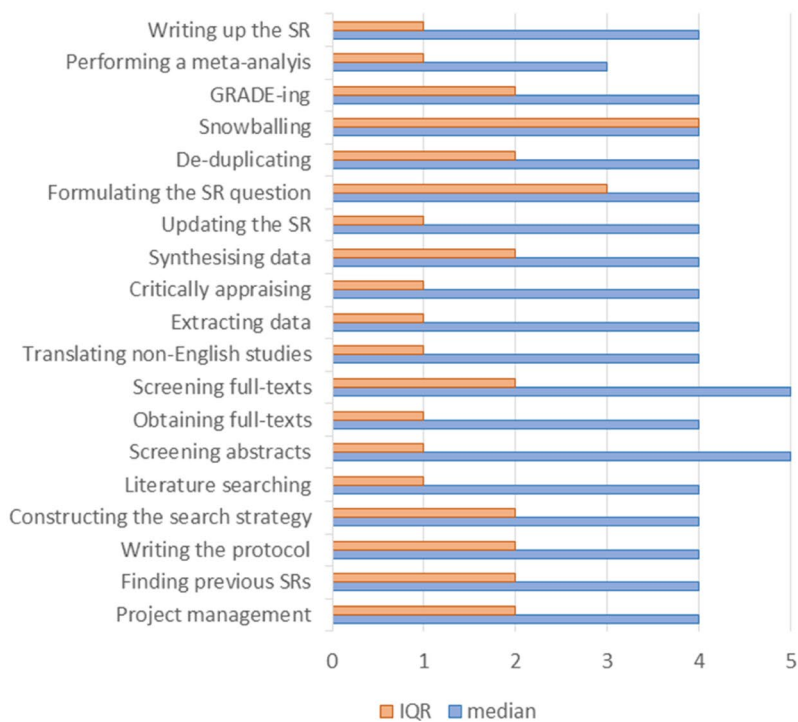
The participants suggested various tools and approaches that need to be developed to increase the efficiency of SRs. Some participants suggested the need for repositories for all registered and published reviews



**Fig. 1** Demographic characteristics of participants. Since there were no extreme outlier values, mean and standard deviation measures were selected as central tendency and level of dispersion measures.\*if respondent's answer was "more than x", "around x" etc. the value was calculated as x

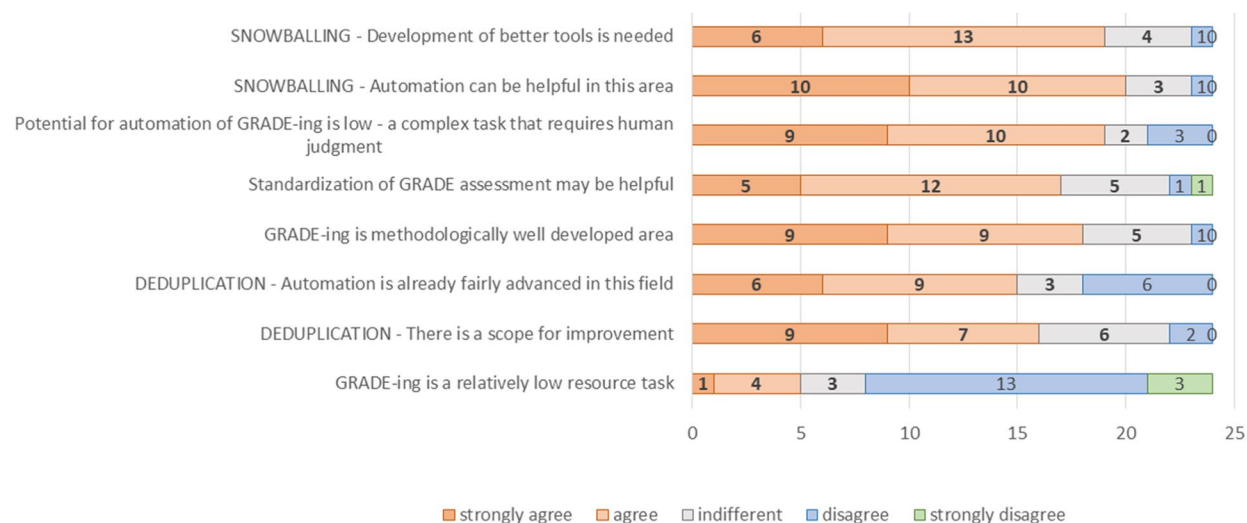


a

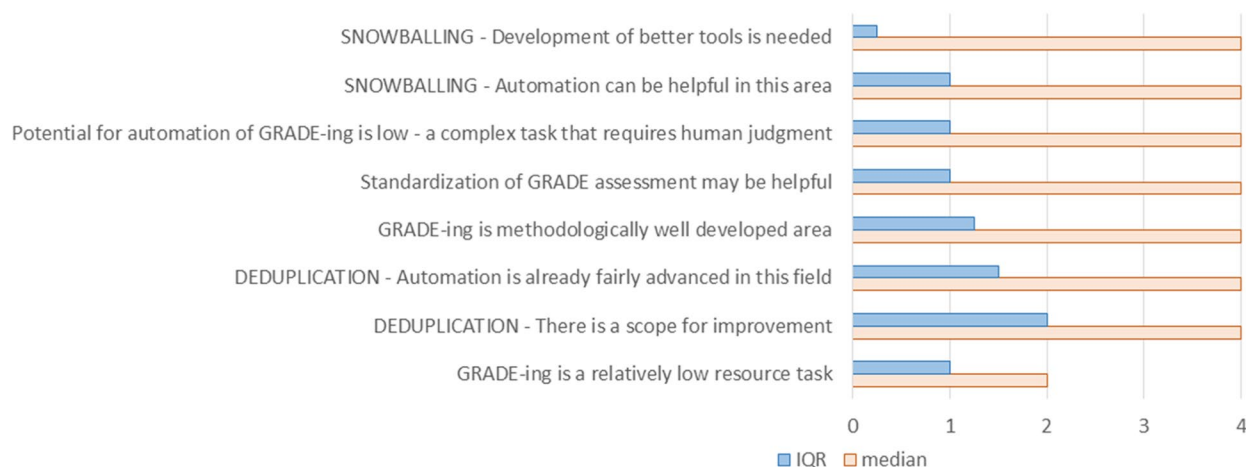


b

**Fig. 2** Results of round 1 of the DS. Participants' view on to what degree the step of the SR production and updating should be prioritized concerning methods of development and automation



a



b

**Fig. 3** Results of round 2 of the DS. Participants’ view on snowballing (development on better tools, need for automation), GRADE-ing (complex task that requires human judgement, standardization, resource use, methodologically developed area), and deduplication (scope for improvement, advancement of automation)

that can be used to find previous SRs and share translations or extracted data across reviews.

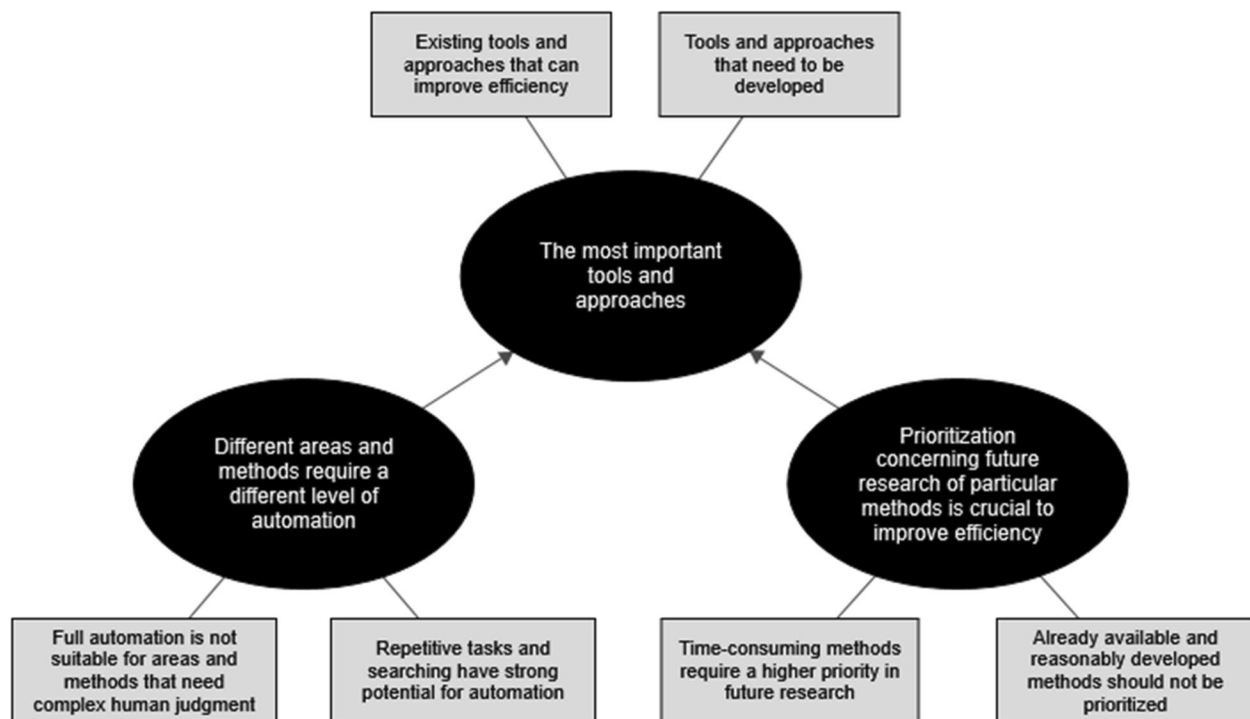
One possible approach to improving SRs’ efficacy would be a common search strategy for all databases. Automation of screening abstracts and full texts is considered one of the most needed tools in SRs. Development of tools for automation in extracting data could be highly beneficial, so future research in that area is worthwhile. Some of the participants also mentioned the need for further development of machine learning to improve the efficiency of SRs, especially for finding full texts.

Development of more advanced tools for the translation of non-English studies was proposed by few participants.

Automation was one of the most frequently mentioned topics in round 1 of our DS. Some participants suggested that particular areas and methods of SRs have a limited potential for automation, and that some parts of the SR process are impossible to automate.

On the other hand, some participants mentioned areas and methods with a strong potential for automation. The components most frequently mentioned with a strong potential for automation are searching and screening.

Since the participants most frequently described some areas and methods as more suitable for automation than others, two additional sub-themes were also developed: “Full automation is not suitable for areas



**Fig. 4** The most promising areas and methods to improve the efficient production of SRs thematic map. The reflexive thematic analysis of qualitative data identified 3 main topics and 6 subtopics

and methods that need complex human judgment” and “Repetitive tasks and searching have strong potential for automation”.

Some participants believed that the components of SR that require complex human judgments are not suitable for automation. One participant emphasized that automation of judgments should not be favourable even if artificial intelligence would allow it.

The participants mostly viewed areas and methods connected with writing as less suitable for automation, specifically the writing up the protocol or manuscript and formulating the review question.

Although some experts considered activities connected with searching as most promising for automation, others suggested that humans still need to make decisions in those areas.

Screening full texts was specified as a part of searching that is not suitable for automation, especially in social sciences where inconsistency on how studies are reported is common.

GRADE-ing or going from evidence to decision is another process that was seen as dependent upon human judgements.

The participants also pointed out that synthesizing data is a complex process, which is why they remain sceptical about its automation.

However, other participants suggested that some areas and methods that need complex human judgments could benefit from semi-automation in which a combination of automation and human judgment will be utilized in the decision-making process.

The participants discussed areas and methods in which automation could be worthwhile. They concluded that repetitive tasks and activities connected with searching have the most substantial potential for automation.

Some participants mentioned in their comments that some areas are very time-consuming, so additional research and development of new tools would be beneficial for improving the efficiency of SR.

Others pointed out that some areas and methods are already developed or automated, which is why they are less important for future research.

Therefore, a third theme emerged: prioritization concerning future research of particular methods to improve the efficiency of SRs in general. Two sub-themes were developed within this theme: “Time-consuming methods require a higher priority in future research” and “Already available and reasonably developed methods should not be prioritized”.

Most participants believed that the SR process stands to benefit the most from future research aimed at improving the time-consuming components.



**Table 2** Qualitative analysis—themes, subthemes, and participants’ statements

Theme	Subthemes	Participants’ statements
The most important tools and methodological approaches	Existing tools and approaches that can improve efficiency	<p>P9: Already methods (SUAMRI, What review is right for you, IEBHC Review Wizard)</p> <p>P2: Revman and most other organizations give templates to fill in details</p> <p>P16: [...] but there are a number of excellent tools being developed for this at the Institute for Evidence-Based Healthcare at Bond University in Australia—<a href="https://fehb.bond.edu.au/education-services/research-tools">https://fehb.bond.edu.au/education-services/research-tools</a></p> <p>P16: I know there are tools for this [De-duplicating], either standalone (e.g., Institute for Evidence-Based Practice at Bond University) or integrated into systematic review software—Covidence, EPPI-Reviewer</p> <p>P11: [Screening full-texts] Already explored in some commercial platforms (i.e., Distiller, EPPI-Reviewer)</p> <p>P16: This [Screening full-texts] is harder to automate at the moment, and is quite a manual task, so a productive area for research</p> <p>P7: Using a program like Covidence or DistillerSR or EPPI-Reviewer to screen the search, perform a risk of bias, extract data—such a program will make sure that all parts of the process are accounted for</p> <p>P32: this [Snowballing] is important but great tools like Citation chaser already exist</p> <p>P7: Use Google Translate whenever there is a non-English publication—Google Translate only trustworthy from any language to English</p> <p>P11: Several platforms for translation are already available (i.e., DeepL)</p> <p>P16: Google Translate as a tool has increased dramatically in quality over the years, and guidance now recommends it can be used for study screening, if not final full-text translation for incorporation into the review</p> <p>P14: Prioritisation of teaching how to use systematic review automation tools in training (e.g., JBI, Cochrane, other organisations)</p> <p>P21: Regular and in-depth training is important</p> <p>P14: Prioritisation of teaching how to use systematic review automation tools in training (e.g., JBI, Cochrane, other organisations)</p> <p>P2: Finding previous systematic reviews is straightforward (provided they are published or registered) hence I believe not much development and automation is required. However, if there is a platform wherein all the registered and published reviews are pulled together like a repository it may benefit researchers</p> <p>P16: Repositories for sharing translations might also be useful, reducing resource use through duplication of effort</p> <p>P16: I have also heard discussions about possible platforms for sharing extracted data across reviews, which could reduce duplication</p> <p>P2: Also, rather than searching in various databases separately, each database can be linked to having one common search strategy, which I believe will save a lot of time</p> <p>P32: One of the most time-consuming parts of a review is wading through irrelevant records made necessary by inconsistent indexing in databases and different syntax in different databases. Findings a way to run a search in one place that would automatically search all sources would be wonderful</p> <p>P8: Machine learning is helping but even more could be done</p> <p>P16: [Screening full-texts] Tools which make this easier, including summarising the results/reasons (e.g., Covidence, EPPI-Reviewer) are already useful</p> <p>P27: This [Screening abstracts] is very tedious and time-consuming task. Artificial intelligence or citizen science might help with that</p> <p>P24: [Extracting data] More challenging to automate, but high value if it can be made to work</p> <p>P13: I think that there are many nice tools already. But two areas which are very time consuming and would benefit from new tools (in my opinion) are finding full text and extracting data</p> <p>P2: Many researchers have to restrict the search to English languages because of resource constraints. If this area is developed, it will be very good</p> <p>P12: this [Translating non-English studies] would be a game changer. ...if... it went beyond quantitative data, and qualitative researchers could be confident the ‘meaning’ of text was captured</p>
	Tools and approaches that need to be developed	

**Table 2** (continued)

Theme	Subthemes	Participants' statements
Different areas and methods require a different level of automation	General	<p>P24: I think the potential for automation to help is limited here [Project management]</p> <p>P11: I am not sure if this step [Formulating the review question] could be automated</p> <p>P27: Anything related to search obviously has a potential for automation, therefore worth researching</p> <p>P12: [Screening abstracts] potentially the strongest area of contribution for smart automation</p> <p>P32: The most time-consuming parts of the process are those that lend themselves best to some degree of automation—deduplication, screening, data extraction. The other parts, I feel, require expert human input especially where complex decisions need to be made</p> <p>P22: [...] the systematic review process is a multi-step process, nearly all of which require judgments. And you cannot automate judgments, and perhaps should not, even in the face of artificial intelligence (AI)</p> <p>P6: [Formulating the review question] this is a piece of the endeavour that I can't see being done for a machine. Whether it is theoretically possible to find ways around this limitation, I doubt research on automation in this area will provide many things</p> <p>P5: [Writing the protocol] Some parts of the protocol follow rules and standards, that can be supported by software and text template, while other elements (e.g. study selection criteria, inclusion of non-randomized evidence) are complex questions, which have to be left to human minds</p> <p>P5: [Writing up the review manuscript] Automatic links between text and key statistical findings are useful, but writing the introduction and discussion section requires a human mind</p> <p>P32: [Constructing the search strategy] I don't know that automation will improve things, a more pressing issue is SRs conducted without input from an info specialist</p> <p>P7: [Literature searching] Difficult to automate, always needs a human to make decisions, but crucial for a systematic review</p> <p>P7: [Screening full-texts] I am afraid that human decisions are needed</p> <p>P32: I am sceptical on automation here [Screening full-texts] especially in the social sciences where there is so much inconsistency on how studies are reported that it takes a human to find the relevant information</p> <p>P7: [GRADE-ing] This process is dependent upon judgements that I can't imagine a meaningful automation or semi-automation</p> <p>P8: [GRADE-ing] I do not think that GRADE decisions should be automatic as they require a lot of reflection and thought. There are already existing tools that are functional</p> <p>P16: [Synthesizing data] I think this is a stage of the review where it's really productive and important for authors to spend time and energy to do this thoughtfully and appropriately—at this stage I think there are risks in trying to automate this, as judgement is needed before proceeding to synthesise data</p> <p>P32: [Synthesizing data] with more complex data sets having a human checking statistics and conversions of different measures to a common effect estimate takes real skill and knowledge of stats so I am sceptical about automating that process</p> <p>P7: [Extracting data] My experience is that it is very demanding and includes a lot of decisions by humans, but a combination would be good</p> <p>P3: [Extracting data] Semi automation perhaps?</p>
Full automation is not suitable for areas and methods that need complex human judgment		<p>P4: Researchers need to understand that automation should prevent them from having to do rote, complicated, repetitive tasks—thus freeing them up to do more interesting and critical tasks. I.e. automation is a tool for them to have more of a difference, whether in evidence-based medicine or policy. It is not a replacement for them</p> <p>P17: Searching and analysing relevance are most likely places for automatization, perhaps also the data extraction</p> <p>P23: AI helps reducing the same work</p> <p>P27: Anything related to search obviously has a potential for automation, therefore worth researching. Especially in the area in qualitative evidence synthesis we still have a lot to research and learn</p>
Repetitive tasks and searching have strong potential for automation		

**Table 2** (continued)

Theme	Subthemes	Participants' statements
Prioritization concerning future research of particular methods is crucial to improve efficiency	General	P13: <i>think that there are many nice tools already. But two areas which are very time consuming and would benefit from new tools (in my opinion) are finding full text and extracting data</i> P3: <i>[De-duplicating] Less important, as already fairly advanced</i> P27: <i>[De-duplicating] Isn't that automated already?</i> P32: <i>[Literature searching] One of the most time-consuming parts of a review is wading through irrelevant records made necessary by inconsistent indexing in databases and different syntax in different databases. Findings a way to run a search in one place that would automatically search all sources would be wonderful</i> P2: <i>[Obtaining full-text] This step takes a lot of time and thus to save time prioritization in future research is needed</i> P13: <i>[Extracting data] This is the most time consuming part where to my knowledge there are no good tool available</i> P16: <i>[Extracting data] This is a very time consuming process and existing tools can be challenging to use, especially for complex reviews with multi-component, highly variable interventions and a lot of variability in how outcomes are measured and reported. Ongoing improvement of data extraction tools would be great, as would semi-automation to assist in identifying and classifying relevant information and reduce author workload. I have also heard discussions about possible platforms for sharing extracted data across reviews, which could reduce duplication</i>
	Time-consuming methods require a higher priority in future research	
Open-ended comments	General	P11: <i>[Obtaining full-text] Already available in most of the commercial software for management of references</i> P6: <i>Whether that [Translating non-English studies] could be extremely valuable in reviews, a lot of research and experimenting is already been done in this area, so prioritizing it in the context of reviews seems unlikely to produce additional benefits</i> P10: <i>"We should develop methods to combine different study designs and generate evidence."</i> P10: <i>"Non-English databases are usually not included in the systematic reviews. For example, it is difficult to get access or translation facilities for Chinese databases. Thereby we are missing a huge chunk of information that could have an impact on the results of the systematic review."</i> P24: <i>"You cannot automate judgements, and perhaps should not, even in the face of artificial intelligence (AI)"</i> P24: <i>"In a bit more 'Intellectual' activities like Rob, synthesis and GRADE-ing I am sceptic towards the automation"</i> P24: <i>"I think one of the challenges when thinking about automation is that people tend to think (as in the case in the survey) of automation supplement/replacing/assisting in existing human processes."</i> P22: <i>"The theoretical framework regarding the SR methodology is clear and valid. Available research shows inconsistent judgements on the risk of bias, methods around data synthesis that are not always appropriate, and we may question the assessment of certainty of evidence... We should start with this; identifying areas of SR methodology that have shown to be inconsistent."</i>
	Already available and reasonably developed methods	

Extracting data was seen as one of the most time-consuming parts of SR, so the participants especially emphasised prioritising that area in future research.

The participants mentioned several methods that are reasonably developed and already available on the market.

They emphasized that already existing methods and areas should not be prioritized in future research since it seems unlikely to produce additional benefits for SR efficiency.

One participant mentioned that there is a need to increase the integration of different study designs into producing evidence.

Although many participants stated that there are good enough translating tools available, the language barrier still resembles a huge obstacle in SRs' production.

Many participants agreed that there is a lot of automation achieved in the area of SRs' production and updating, but there are still tasks that cannot and should not be automatized since they rely on complex human judgements.

Several participants emphasized that methodology of SRs' production is fairly advanced.

## Discussion

The main finding of this DS was that extracting data, literature searching, and screening abstracts as the most important areas to be prioritized in future research when developing SR methods and tools.

There is a consensus among participants that "snowballing" is a relatively low resource task, development of better tools is needed, and automation can be helpful in this area. One participant (P32) mentioned an efficient tool that has a great potential to successfully automate this step: *Citationchaser*. There are some software solutions already available that support basic forms of snowballing/citation chasing. *Citationchaser* [26] is an open source, easy-to-use tool for quick backward and forward citation chasing, developed by Haddaway et al. [27], and seems to be the most advanced in the field. It generates standardized output files that can be quickly and effectively combined with the results of bibliographic database searches to reduce duplication and increase the breadth of the pool of potentially pertinent records that can be screened within an evidence synthesis [27]. The fact that only one participant was aware of existing of this tool, among our highly expert panel, grants that this tool has to be further developed and popularized.

Participants also agreed that potential for automation of "GRADE-ing" is low since this is a complex task that requires human judgement, and that methodologically this is a very well-developed area and standardization of GRADE assessments may be helpful. Regarding

"deduplication," experts agreed that automation is already fairly advanced, but there is room for improvement.

Regarding possible area of improvements in methodology, several participants emphasized that automation is not a panacea and has to be used to "prevent from having to do rote, complicated repetitive tasks"; "automation is a tool...to have more difference...it's is not a replacement for human judgement".

Including non-English studies in SRs has been recognized as important to avoid bias, although reviewers commonly report that it is costly and time-consuming to include them, and previously have been reluctant to bother with the language barriers [28]. "Many researchers have to restrict the search to English language because of resource constraints". Participants from our DS showed awareness and willingness to incorporate non-English evidence in their SRs. In their comments, participants emphasized the importance of this issue, especially in the qualitative area: "This would be a game changer...qualitative researchers could be confident the 'meaning' of text was captured". Many participants emphasized that quality of available language translating tools has dramatically increased over the years, specifically pointing out Google translate and Deep L, which is promising, and hopefully will progress into qualitative field of research in the future.

## Limitations

Limitations of the study stem from the very nature of the research method: one can always debate that there were additional "expert" SRs' authors who could have better answered the survey. Efforts were made to select experts who were relatively impartial yet had interest in the research topic and were willing to spare their precious time. The EVBRES collective knowledge of the SRs' production landscape was excellent base for handpicking the best available sample and serve as effective gatekeepers [18]. In fact, participants demonstrated vast experience (totalling 440 years in conducting SRs) at a relatively young age (most panellists were 41–50 years of age) (Fig. 1). Most of the participants (26/33) work in academia: it is highly understandable that researchers from that area are the most efficient producers of SRs. Participants published SRs in various settings: all ( $n=33$ ) have published in various journals, and the majority ( $n=28$ ) published Cochrane SRs. Another limitation is also due to the nature of DSs: the Delphi method has been criticized in that it does not allow participants to discuss the issues raised and gives no opportunities for participants to elaborate on their views, resulting in the potential risk that greater reliance is placed on the results than might be warranted [18]. DSs also have additional limitations, such as not allowing the same level of interaction or fast

turnaround that is possible, for example, in a focus group. However, this also presents a strength due to the fact that participants do not meet with each other face to face, and therefore they can present and react to ideas unbiased by the identities and pressures of others [29].

## Conclusions

The participants recommended tools and approaches that can improve the efficiency of SRs. Data extraction was prioritized by the majority of participants as an area that needs more research/methods development, where research and automation can help reduce the intensity of resource use. They specified that some areas and methods are more suitable for automation than others, e.g., snowballing, and development of tools is needed in this area. There is an open-source tool—Citation chaser, which has a high potential to present a significant time saving in the SRs production process. GRADE-ing was identified as an area that is methodologically well developed, a complex task that has lowest potential for automation, as it requires high level of human judgement.

As expected, s of SR automation is already developed and less critical for future research (GRADE-ing), om additional research and the development of new tools.

## Abbreviations

SR	Systematic review
DS	Delphi study
JBI	Joanna Briggs Institute
WG	Working group
COST	European Cooperation in Science and Technology
EVBRES	EVidence-Based REsearch

## Acknowledgements

We kindly thank EVBRES members Hans Lund, Mónica Sousa, James Thomas, Kontogianni Meropi, and María E. Marqués for useful comments on the protocol and help with recruiting participants. We also thank Affan Kaknjo for technical support and Eliana Ben Sheleg for critically reading the manuscript.

## Data protection, privacy, and ethics

This study posed negligible risk of discovering sensitive information. All steps were taken to minimize this risk by asking all participants to not provide names of people/institutions in the survey and by (pseudo) anonymization of the data before sending it to others.

No identifying information was collected at any stage of the DS. The online tool was set not to collect IP addresses, and participants were identified exclusively by a study ID number. The ID number was associated with the participants' email address which was used to match participants in the two rounds of the DS and was stored in a password-protected computer file available only to the principal investigator. All data collected remain confidential, since privacy was the highest priority in gathering, storing, and handling data. All data were converted to an appropriate MS Office format and will be stored on the personal data storage device of the principal investigator for 5 years, password-protected and coded.

## Authors' contributions

MMK: conceptualization, methodology, investigation, formal analysis, writing—original draft, writing—review and editing. VT: methodology, investigation, formal analysis. MEE: conceptualization, methodology, writing—review and editing. BNS: conceptualization, methodology, writing—review and

editing. RS: conceptualization, methodology, writing—review and editing. EB: conceptualization, methodology, writing—review and editing. NR: conceptualization, methodology, writing—review and editing. APK: conceptualization, methodology, writing—review and editing. AM: conceptualization, methodology, formal analysis, writing—review and editing. All authors read and approved the final manuscript.

## Funding

This work was partly supported by resources from the EU funded COST Action EVBRES (CA17117). The funder had no role in the design of this study, during its execution and data interpretation.

## Availability of data and materials

Anonymized datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

## Declarations

### Ethics approval and consent to participate

Ethical approval for this study was granted from the Human Subjects Research Committee of Ben-Gurion University of the Negev, P.O.B. 653, Beer Sheva, Israel, under the number ME21042021, on April 21st, 2021.

### Consent for publication

At the beginning of the survey, there was an introductory text explaining details of the DS and a box that needed to be ticked by the participant to provide consent before the participant could proceed with the survey. Participants were given opportunity at each e-mail invitation to unsubscribe from further notices and leave the study at any time, without explanation, if they did not want to participate in the DS further. Participants were informed that all their answers are anonymized.

### Competing interests

None of the authors report any financial conflicts of interest with respect to the topic of this manuscript. All authors have a general interest in evidence synthesis methods. Some authors are associated with groups, conferences, and tools focusing on evidence synthesis methods: Barbara Nussbaumer-Streit is co-convenor of the Cochrane Rapid Reviews Methods Group. Raluca Sfetcu is a member of the JBI method group for "Systematic reviews of etiology and risk". Ana Marušić is funded by the Croatian Science Foundation under Grant agreement No. IP-2019-04-4882.

### Author details

<sup>1</sup>Cantonal Hospital Zenica, Crkvice 67, 72000 Zenica, Bosnia and Herzegovina. <sup>2</sup>Sarajevo Medical School, Sarajevo School of Science and Technology, Hrasnička Cesta 3a, 71210 Ilidža, Bosnia and Herzegovina. <sup>3</sup>ST-OPEN, University of Split School of Medicine, Split, Croatia. <sup>4</sup>Department of Research in Biomedicine and Health, Center for Evidence-Based Medicine, University of Split School of Medicine, Split, Croatia. <sup>5</sup>Department of Health Policy and Management, Guilford Glazer Faculty of Business and Management and Faculty of Health Sciences, Ben-Gurion University of the Negev, Beersheba, Israel. <sup>6</sup>Institute of Health Policy Management and Evaluation, Dalla Lana School of Public Health, University of Toronto, Toronto, Canada. <sup>7</sup>Cochrane Austria, Danube University Krems, Krems a.d. Donau, Austria. <sup>8</sup>Department of Psychology, Spiru Haret University, Bucharest, Romania. <sup>9</sup>National School of Public Health, Management and Professional Development Bucharest, Bucharest, Romania. <sup>10</sup>Centro de Análisis de la Evidencia Científica, Academia Española de Nutrición y Dietética, Pamplona, España. <sup>11</sup>Department of Pathology, Faculty of Medicine and Surgery, University of Malta, Msida, Malta. <sup>12</sup>Department of Applied Health Research, University College London, London, UK. <sup>13</sup>Department of Nursing, Cyprus University of Technology, Limassol, Cyprus.

Received: 8 September 2022 Accepted: 20 March 2023

Published online: 28 March 2023

## References

1. Bastian H, Glasziou P, Chalmers I. Seventy-five trials and eleven systematic reviews a day: how will we ever keep up? *PLoS Med.* 2010;7(9):e1000326.

2. Chalmers I, Hedges LV, Cooper H. A brief history of research synthesis. *Eval Health Prof.* 2002;25(1):12–37.
3. Brunton G, Stansfield C, Caird J, Thomas J. Finding relevant studies. In: Gough D, Oliver S, Thomas J, editors. *An introduction to systematic reviews*. 2nd ed. London: Sage Publications Inc; 2017.
4. Thane P. A critical woman. Barbara Wootton, social science and public policy in the twentieth century. Vol. 23. By Ann Oakley. *Twentieth Century British History*; 2012.
5. Gopalakrishnan S, Ganeshkumar P. Systematic reviews and meta-analysis: Understanding the best evidence in primary healthcare. *J Fam Med Prim Care.* 2013;2(1):9. Available from: [www.jfmpc.com](http://www.jfmpc.com).
6. Uman LS. Information management for the busy practitioner systematic reviews and meta-analyses information management for the busy practitioner. *J Can Acad Child Adolesc Psychiatry.* 2011;20(1):57. Available from: [www.cochrane.org](http://www.cochrane.org).
7. Petrosino A. Reflections on the genesis of the Campbell Collaboration. *The Experimental Criminologist.* 2013. p. 9–12. Cited 2022 Feb 8. Available from: [moz-extension://2adc957f-a2f0-4775-8ddc-2286d82c793a/enhanced-reader.html?openApp&pdf=https%3A%2F%2Fcampbellcollaboration.org%2Fimages%2Fpdf%2Fplain-language%2FPetrosino\\_2013\\_EC\\_Reflections\\_Genesis\\_of\\_the\\_Campbell\\_Collaboration.pdf](https://doi.org/10.1186/s13643-018-0786-6).
8. Littell JH, White H. The Campbell Collaboration: providing better evidence for a better world. *Res Soc Work Pract.* 2018;28(1):6–12.
9. Munn Z, Peters MDJ, Stern C, Tufanaru C, McArthur A, Aromataris E. Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. *BMC Med Res Methodol.* 2018;18(1):1–7.
10. Nussbaumer-Streit B, Ellen M, Klerings I, Sfetcu R, Riva N, Mahmić-Kaknjó M, et al. Resource use during systematic review production varies widely: a scoping review. *J Clin Epidemiol.* 2021;139:287–96.
11. Nussbaumer-Streit B, Ziganshina L, Mahmić-Kaknjó M, Gartlehner G, Sfetcu R, Lund H. Resource use during systematic review production varies widely: a scoping review: authors' reply. *J Clin Epidemiol.* 2022;142:321–2.
12. EVBRES. Evidence based research. Available from: <https://evbres.eu>. Cited 2022 Jul 13.
13. Ellen M, Sfetcu R, Baladia E, Nussbaumer-Streit B. Why conducting and updating systematic reviews are resource intensive: a phenomenological qualitative study. G. Balint, Antala B, Carty C, Mabieme J-MA, Amar IB, Kaplanova A, editors. *Manag Heal.* 2019;23(4):8–11. Available from: <https://cris.bgu.ac.il/en/publications/why-conducting-and-updating-systematic-reviews-are-resource-inten>. Cited 2022 Feb 9.
14. Lisa A, Lotty H, Jos K, Gerald G, Barbara N-S, Mersiha M-K, et al. Improving efficiency of systematic reviews production through an exploration of available methods and tools – a scoping review. Available from: <https://osf.io/9423z/>. Cited 2022 Jul 13.
15. Barrett D, Heale R. What are Delphi studies? *Evidence-Based Nursing.* 2020;23(3):68–9.
16. Joyner HS, Smith D. Using Delphi surveying techniques to gather input from non-academics for development of a modern dairy manufacturing curriculum. *J Food Sci Educ.* 2015;14(3):88–115.
17. Jünger S, Payne SA, Brine J, Radbruch L, Brearley SG. Guidance on conducting and reporting Delphi studies (CREDES) in palliative care: recommendations based on a methodological systematic review. *Palliat Med.* 2017;31(8):684–706. [cited 2022 Feb 9]. Available from: <https://pubmed.ncbi.nlm.nih.gov/28190381/>.
18. Hasson F, Keeney S, McKenna H. Research guidelines for the Delphi survey technique. *J Adv Nurs.* 2000;32(4):1008–15.
19. Ellen M, Sfetcu R, Baladia E, Nussbaumer-Streit B. Why conducting and updating systematic reviews are resource intensive: a phenomenological qualitative study protocol. *Manag Heal.* 2020;23(4):8–11.
20. Akins RB, Tolson H, Cole BR. Stability of response characteristics of a Delphi panel: application of bootstrap data expansion. *BMC Med Res Methodol.* 2005;5(1):1–2.
21. Okoli C, Pawlowski SD. The Delphi method as a research tool: an example, design considerations and applications. *Inf Manag.* 2004;42(1):15–29.
22. Google Scholar. Available from: <https://scholar.google.com>. Cited 2022 Jul 13.
23. Tsafnat G, Glasziou P, Keen Choong M, Dunn A, Galgani F, Coiera E. Systematic review automation technologies. 2014. Available from: <http://www.systematicreviewsjournal.com/content/3/1/74>
24. Braun V, Clarke V, Hayfield N, Terry G. Thematic analysis BT - handbook of research methods in health social sciences. Springer Nat Singapore Pte Ltd. 2019;
25. Braun V, Clarke V. Using thematic analysis in psychology. *Qual Res Psychol.* 2006;3(2):77–101.
26. Haddaway NR, Grainger MJ, Gray CT. Citationchaser: an R package and Shiny app for forward and backward citations chasing in academic searching. Available from: <https://estech.shinyapps.io/citationchaser/>. Cited 2022 Jul 20.
27. Haddaway NR, Grainger MJ, Gray CT. Citationchaser: a tool for transparent and efficient forward and backward citation chasing in systematic searching. *Res Synth Methods.* 2022;13:533. Available from: <https://guides.library.illinois.edu/c.php?g=>
28. Neimann Rasmussen L, Montgomery P. The prevalence of and factors associated with inclusion of non-English language studies in Campbell systematic reviews: a survey and meta-epidemiological study. <https://doi.org/10.1186/s13643-018-0786-6>. Cited 2022 Feb 9
29. Diamond IR, Grant RC, Feldman BM, Pencharz PB, Ling SC, Moore AM, et al. Defining consensus: a systematic review recommends methodologic criteria for reporting of Delphi studies. *J Clin Epidemiol.* 2014;67(4):401–9.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

### Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more [biomedcentral.com/submissions](https://biomedcentral.com/submissions)

