BMJ Open

Exploring decision-makers' challenges and strategies when selecting multiple systematic reviews: insights for AI decision support tools in healthcare

Carole Lunny ,[1,2] Sera Whitelaw,[3] Emma K Reid ,[4] Yuan Chi ,[5] Nicola Ferri ,[6] Jia He (Janet) Zhang,[7] Dawid Pieper,[8] Salmaan Kanji,[9,10] Areti-Angeliki Veroniki,[11,12] Beverley Shea,[13] Jasmeen Dourka,[14] Clare Ardern,[15] Ba Pham,[16] Ebrahim Bagheri,[17] Andrea C Tricco [12,18]

To cite: Lunny C, Whitelaw S, Reid EK, *et al*. Exploring decision-makers' challenges and strategies when selecting multiple systematic reviews: insights for AI decision support tools in healthcare. *BMJ Open* 2024;**14**:e084124. doi:10.1136/bmjopen-2024-084124

► Prepublication history and additional supplemental material for this paper are available online. To view these files, please visit the journal online (https://doi.org/10.1136/bmjopen-2024-084124).

Received 10 January 2024
Accepted 24 June 2024

For numbered affiliations see end of article.

Correspondence to
Dr Carole Lunny;
carole.lunny@ubc.ca

ABSTRACT

Background Systematic reviews (SRs) are being published at an accelerated rate. Decision-makers may struggle with comparing and choosing between multiple SRs on the same topic. We aimed to understand how healthcare decision-makers (eg, practitioners, policymakers, researchers) use SRs to inform decision-making and to explore the potential role of a proposed artificial intelligence (AI) tool to assist in critical appraisal and choosing among SRs.

Methods We developed a survey with 21 open and closed questions. We followed a knowledge translation plan to disseminate the survey through social media and professional networks.

Results Our survey response rate was lower than expected (7.9% of distributed emails). Of the 684 respondents, 58.2% identified as researchers, 37.1% as practitioners, 19.2% as students and 13.5% as policymakers. Respondents frequently sought out SRs (97.1%) as a source of evidence to inform decision-making. They frequently (97.9%) found more than one SR on a given topic of interest to them. Just over half (50.8%) struggled to choose the most trustworthy SR among multiple. These difficulties related to lack of time (55.2%), or difficulties comparing due to varying methodological quality of SRs (54.2%), differences in results and conclusions (49.7%) or variation in the included studies (44.6%). Respondents compared SRs based on the relevance to their question of interest, methodological quality, and recency of the SR search. Most respondents (87.0%) were interested in an AI tool to help appraise and compare SRs.

Conclusions Given the identified barriers of using SR evidence, an AI tool to facilitate comparison of the relevance of SRs, the search and methodological quality, could help users efficiently choose among SRs and make healthcare decisions.

STRENGTHS AND LIMITATIONS OF THIS STUDY

⇒ Our study was conducted in accordance with our protocol published a priori.
⇒ We attempted to maximise survey reach and responses by using emails and social media distribution. However, our email response rate was low (7.9%) and due to the nature of social media advertising, we were unable to calculate a true survey response rate.
⇒ Our targeted emails and social media advertising may have missed important decision-makers that use systematic reviews. Individuals involved in guideline development and policymaking may have been more likely to respond to the survey.
⇒ Response bias is a factor as survey respondents working in higher income countries were over-represented.

BACKGROUND

Evidence-informed healthcare requires decision-makers to identify evidence, appraise its methodological quality and relevance and apply it to a particular practice scenario.[1] Systematic reviews (SRs) are used to collate evidence from primary studies (eg, randomised controlled trials, cohort studies), which are appraised and synthesised using systematic methods.[2] Methodologically sound SRs that are well reported and with a low risk of bias are widely regarded as a gold standard for healthcare decision-making.[3–5]

Each year, there is an exponential rise in the volume of research produced in healthcare,[6] including SRs.[7–9] Between 2000 and 2019, the number of SRs produced annually increased 20-fold, with 80 SRs published per day by 2019.[10] A surge of SR publications during the COVID-19 pandemic has also been observed.[11–13] Along with the increasing prevalence of SRs overall, there has been an increase in the number of duplicated SRs with the same or similar research questions and eligibility criteria.[14 15] Between 2000 and 2020, approximately 1200 and 1600 systematic

review clusters (ie, duplicated SRs) addressing the same clinical, public health or policy questions were identified by two bibliometric studies.[16 17] Duplicated publications increased over time, with the highest increase occurring in the most recent 5-year period (2016 to 2020).[16 17]

Despite similar research questions and focus, duplicated SRs may report discordant results and conclusions. Research groups have studied overlapping SRs in attempt to examine and identify sources of discordance[12–14] and approaches for managing discordant SR findings.[15 16] In 1998, Jadad *et al*, published an algorithm tool[18] to help users select the 'best evidence' review among multiple discordant SRs.[19–21] To assess reproducibility of the algorithm, our research group performed a replication study that compared the findings of 21 publications that used the Jadad tool to choose one or more SRs as 'best evidence' with our own independent Jadad assessment.[22] In 62% of cases, replication was unsuccessful, and a different, higher methodological quality SR was chosen by our group. Sources of discrepancies included different Population, Intervention, Comparison, Outcome (PICO) eligibility criteria, databases searched, primary studies and/or analysis methods.[22] These studies highlight the expertise required by experienced researchers to manually assess and compare similar SRs that differ across their results and conclusions.

If a healthcare decision is informed by SR evidence of low methodological quality and where inappropriate methods were used, this risks negative patient care outcomes.[8 23–25] An example of misleading results from SRs with meta-analysis is when ivermectin, an antiparasitic medication, was widely promoted across the world for preventing and treating COVID-19.[24 26–28] Many of the SRs on ivermectin for COVID-19 were discordant[24] and varied in methodological quality. One of these meta-analyses, submitted as a preprint (https://www.researchsquare.com/article/rs-100956/v2), has since been withdrawn, whereas a published meta-analysis (https://academic.oup.com/ofid/article/8/11/ofab358/6316214) has been retracted after it was found to include fraudulent data (note that we intentionally omitted a formal reference to these studies). Despite these serious concerns, millions of doses of ivermectin have already been given to treat or prevent COVID-19 globally, potentially risking unnecessary toxicity and depleting supply where it is otherwise indicated. The biased results from these and other poorly designed and reported SRs can mislead decision-making at all levels.[29–31]

At this time, it is unknown whether decision-makers, such as practitioners and policymakers, struggle with comparing and/or choosing SRs when there are multiple on the same topic, and what barriers, if any, are faced. There is also value in learning, which variables or features of SRs are considered most important when comparing multiple on the same topic. Keeping abreast of the latest research on a topic is already a monumental task.[5] Therefore, we proposed that a tool that incorporates artificial intelligence (AI) to help navigate a growing body of literature would be beneficial and could increase efficiencies.

The purpose of this study was to explore the desire for a proposed AI-informed, evidence-based tool to help healthcare decision-makers appraise and choose among SRs for real-world practice. To understand current need, we surveyed decision-makers to determine how they use SRs to inform their decisions, and how they choose the best SR evidence when there are multiple SRs addressing the same clinical question. Responses will also inform a larger project[32] to develop an automated decision support tool to assess the strengths and weakness of multiple SRs on the same topic.

METHODS

Protocol

The study protocol can be found on the Open Science Framework at https://osf.io/nbcta/. The reporting of this survey is in accordance with the Checklist for Reporting Results of Internet E-Surveys (online supplemental appendix A).[33] Important definitions are found in box 1.

Survey design

Our investigative team used a cross-sectional survey design, informed by established approaches for conducting needs assessments and the Dillman approach for conducting online surveys.[34] The survey was conducted using Qualtrics (Qualtrics Labs, Provo, Utah).[35 36] No incentive or compensation was offered to respondents. Survey responses were anonymous. Personal identifying information was only collected on a voluntary basis from respondents who wished to be contacted about the survey's results. Informed consent was implied when participants ticked a consent box on the first survey page.

We created an English-language survey with 21 questions, primarily close-ended in nature (full survey questions in online supplemental appendix B). The survey questions were subdivided into three parts: (a) demographics, (b) experiences and barriers to choosing SRs when more than one exists on the same topic and (c) data elements to consider when choosing the SRs from multiple on the same topic. Respondents were allowed to skip questions they did not wish to answer and were able to review and change their answers prior to submitting their responses.

We were particularly interested in gaining insight into the specific features an SR decision-makers would consider when comparing and selecting among multiple SRs, as these features would be crucial to inform priority areas for an AI tool. We recognised the potential for 'leading' or influencing survey respondents by providing a curated list of SR features our steering group deemed relevant in advance of their independent responses. To protect against this influence, we created two similar questions (Q11 and 12a–c), which were randomly allocated to half of the respondents: the first (Q11) was a multiple-choice

Box 1 Important definitions

Decision-maker

We define 'decision-makers' as individuals who are likely to be able to use research results to make informed decisions about health policies, programmes and/or clinical practices.[44] A decision-maker can be, but is not limited to, a health practitioner, a policymaker, an educator, a healthcare administrator, a community leader or an individual in a health charity, patient group, private sector organisation or media outlet.[44] The following individuals and groups were considered decision-makers:

⇒ Health practitioner (individuals who provide care, eg, nurses, physicians, pharmacists, mental health counsellors, community-based workers).
⇒ Patient, caregiver, family member, member of the public.
⇒ Patient and consumer advocacy organisation representative, community leader.
⇒ Policymaker (government representative, public funding agency representative, healthcare/hospital administrator, Clinical Practice Guideline developer, Health Technology Assessment developer).
⇒ Educator.
⇒ Industry representative (eg, drug/device manufacturers).
⇒ Researcher and/or academic.
⇒ Information scientist/medical librarian.
⇒ Journal editor, publisher, news media.
⇒ Student, trainee, postdoctoral fellow, graduate student/post graduate trainee/undergoing practicum in a clinical programme or focused on health policy or research.

Evidence-informed decision-making

Evidence-informed decision-making stresses that the best available evidence from research should inform decisions as well as other factors such as context, public opinion, equity, feasibility of implementation, affordability, sustainability and acceptability to stakeholders. It is a systematic and transparent approach that applies structured and replicable methods to identify, appraise and make use of evidence across decision-making processes, including for implementation.

Systematic review

A systematic review attempts to collate all study-specific evidence that fits prespecified eligibility criteria to answer a specific research question. It uses explicit, systematic methods that are selected with a view to minimising bias, thus providing more reliable findings from which conclusions can be drawn and decisions can be made.[2]

Meta-analysis

Traditional meta-analysis is a statistical method to combine the results from two or more primary studies (eg, randomised controlled trials, cohort studies), to produce a point estimate of an effect and measures of the precision of that estimate.[2] We also considered meta-analyses with other quantitative results (eg, meta-analyses of prevalence data).

Network meta-analysis

'Any set of studies that link three or more interventions via direct comparisons forms a network of interventions. In a network of interventions, there can be multiple ways to make indirect comparisons between the interventions. These are comparisons that have not been made directly within studies, and they can be estimated using mathematical combinations of the direct intervention effect estimates available.[2]' A network is composed by at least three nodes (interventions or comparators) and these are connected (graphically depicted as lines/edges) when at least one study compares the underlying two interventions—the direct comparisons. Reviews that intend to compare multiple treatments with a network meta-analysis (NMA) but then find that the expectations or assumptions are violated (eg, the network is 'disconnected', studies are too heterogeneous to combine, underlying assumptions of the method are not met), and hence an NMA is not possible or optimal and are also considered in our definition.[45]

Continued

Box 1 Continued

Discordance

Discordance is when SRs with similar public health, or policy eligibility criteria (as expressed in Population, Intervention, Comparison, Outcome) report different results or conclusions for the same outcome. We define discordant results as differences based on the methodological decisions SR authors make, or different interpretations or judgements about these results.[22]

Risk of bias assessment in the systematic review level

A risk of bias assessment evaluates limitations in the way in which the results were planned, analysed and presented. If these methods are inappropriate, the validity of the findings can be compromised. Bias may also be introduced when interpreting the results to draw conclusions. Conclusions may include 'spin' (eg, biased mis-representation of the evidence, perhaps to facilitate publication) or (erroneous) mis-interpretation of the evidence.[46] Ideally, potential biases identified in the results of the SR might be acknowledged and addressed appropriately when drawing conclusions.[47] Similarly, a well-conducted SR draws conclusions that are appropriate to the included evidence and, therefore, is free of bias even when the primary studies included in the review have high risk of bias. On the basis of the risk of bias assessment, supported by balanced reporting of SR/meta-analysis (MA) interpretation of findings, relevance of included studies to the SR/MA' question, a final consideration is performed on whether the SR/MA as a whole is at 'low', 'high' or 'unclear' risk of bias.

Reporting comprehensiveness

The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) standard encourages reporting completeness or comprehensiveness when authors write up the results of their SRs prior to publication.[48] Often a PRISMA checklist is required when submitting a systematic review to a peer-reviewed journal for consideration. A review can be well conducted, but poorly reported; or poorly conducted but well reported (even if methods were poor). The Enhancing the QUAlity and Transparency Of health Research Network[49] is an international initiative that seeks to improve the reliability and value of published health research literature by promoting transparent and accurate reporting and wider use of robust reporting guidelines.

Quality of conduct (systematic reviews)

Methodological quality is about how well the research is conducted according to established guidance (eg, Cochrane Handbook,[2] JBI Manual). The Assessing the Methodological Quality of SRs (AMSTAR) measurement tool was designed to appraise the quality of conduct of SRs.[50] AMSTAR has been validated and proven popular as a simple means of assessing the quality of reviews.[51 52] A recently updated version (AMSTAR V.2) was published in 2017.[43]

Certainty of the evidence assessment

An assessment of the certainty of evidence is defined as any of evaluation of the strength of the evidence such as the Grading of Recommendations, Assessment, Development and Evaluations (GRADE) approach,[53] criteria for credibility assessment, and other approaches used to grade the overall body of the evidence. GRADE is a well-established approach to assess the certainty of evidence based on the following criteria: risk of bias of the primary studies, imprecision, indirectness, inconsistency and publication bias. GRADE is designed for assessing the certainty of the evidence deriving from primary studies.

drop-down list of elements to choose from, and the second (Q12a–c), included a short case study summarising the characteristics, methods and results from three similar SRs on the same topic, ultimately asking the respondent to choose which SR(s) they would use to inform a discussion with a patient, and which features of the SR led to their decision (online supplemental appendix B).

Eight academics piloted the survey and modified it iteratively to improve clarity, face validity and content validity.

Sample size

The sample size was calculated to detect mean difference of 50% between two factor levels with 90% power.[37] To detect this difference, a sample of 440 decision-makers was required, assuming an SD of 1.6 points (based on similar surveys[38 39]) and a 5% significance level. We assumed that contacting quadruple (ie, 1760) the number of decision-makers would be sufficient to recruit the required number (assuming a 25% response rate), allowing for failed email addresses and non-response.

Distribution of the survey

We aimed to survey individuals from organisations or institutions, which produce SRs, as well as decision-makers of all types who use SRs.

We developed an email list of SR-producing groups, including Cochrane Multiple Treatments Methods Group, Guidelines International Network, JBI (formerly the Joanna Briggs Institute), Campbell Collaboration, US Agency for Healthcare Research & Quality's Evidence-Based Practice Centre programme, Centre for Reviews and Dissemination, Canadian Agency for Drugs and Technologies in Health, Evidence for Policy and Practice Information and Co-ordinating Centre, Clinical Epidemiology programme at the Ottawa Hospital Research Institute, the Grading of Recommendations, Assessment, Development and Evaluations group). These potential survey participants were sent an email describing the purpose of the study, requesting their participation and providing a link to the survey.

We leveraged the professional contacts from our steering committee and distributed the survey to an additional 28 organisations anonymously (online supplemental appendix C). We also included participants from a UBC Methods Speaker Series on evidence synthesis methods (https://www.ti.ubc.ca/2023/01/10/methods-speaker-series-2023/). In addition, we advertised through the e-newsletters of Knowledge Translation Canada, SPOR (Strategy for Patient-Oriented Research) Evidence Alliance and Therapeutics Initiative. We also contacted professional healthcare organisations (eg, Canadian Association for Physiotherapists) to promote the survey in their newsletters.

A distribution plan was followed to disseminate and advertise the survey. The Dillman approach[34] suggests repeated contact to boost responses, which we followed by sending out three reminders to email recipients, and repeated advertisement through social media outlets. Anonymous links were included in LinkedIn and Twitter posts, which were circulated through targeted Twitter accounts, such as the Knowledge Translation Program, SPOR Evidence Alliance and the Therapeutics Initiative. Tweets were retweeted among followers. We used twitter cards (ie, advertisements with pictures) and targeted hashtags to increase awareness of the survey. We also advertised through two LinkedIn accounts.

Timing

The survey ran from 19 July to 19 August 2022. Qualtrics email reminders were scheduled at 2-week intervals throughout this period to unfinished or non-respondents. We estimated that the survey would take approximately 10 min of a respondent's time.

Patient involvement

Patients or the public were not involved in the design, or conduct, or reporting or dissemination plans of our research.

Data analysis

Prior to data analysis, the responses were transferred from Qualtrics to MS Excel. Questionnaires that were terminated before completion were included in analyses, but those that were entirely blank were excluded. We measured the time respondents took to fill in a questionnaire regardless of whether it was complete.

Descriptive statistics were calculated for each closed response question, including count, frequency, with denominators taken as the number who provided a response to the question. One researcher coded responses to the open-ended questions or comments independently by identifying themes. The free-form responses were then presented descriptively as counts and frequencies in an identical fashion to closed responses. We stratified the analysis by type of decision-maker as presented in box 1. When a respondent indicated that they were more than one type of decision maker, we calculated the response for all their respondent types. For example, if they identified as both a practitioner and researcher, and responded yes to question 1, we counted a yes for both practitioner and researcher types. We compared the results of the two randomly presented questions.

RESULTS

Recruitment results

A total of 3158 email invitations were sent to advertise the survey. Of these, 197 emails failed to reach the recipients due to incorrect addresses, no longer at the related job post, etc, resulting in a total of 2961 email invitations successfully delivered (figure 1). After consolidating duplicates (n=25) and blank responses (n=83), a total of 684 survey responses were included in the analysis. Most respondents completed the survey by clicking and completing an anonymous link distributed over social media and e-newsletters (n=450), compared with those

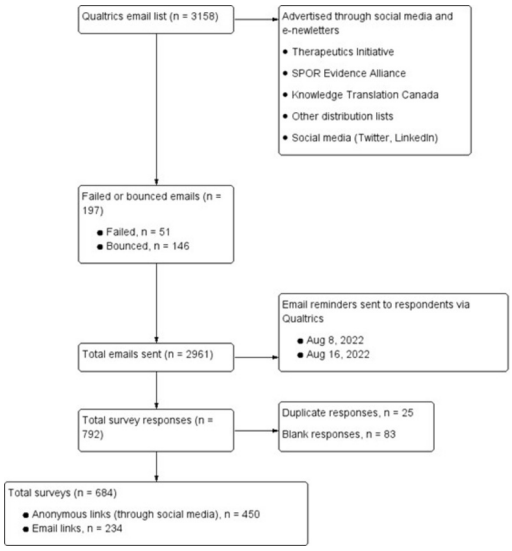

Figure 1 Recruitment of decision-makers. SPOR, Strategy for Patient-Oriented Research.

who responded through the Qualtrics email link (n=234). As per our sample size calculation, we expected a 25% response rate to email invitations but only achieved 7.9%.

Of the 684 respondents, 462 (67.5%) answered all the survey questions, and 97 (14.2%) completed less than 50% of the questions. For those who completed the survey, the median response time was 11 min.

Demographics and characteristics of respondents

The majority of surveyed decision-makers identified as researchers (58.2%), practitioners (37.1%), students/trainees (19.2%) and policymakers (13.5%), and many respondents identified as more than one role. For example, of the 62 (13.5%) policymakers, 37 (59.7%) also identified as a researcher, 14 (22.6%) as a practitioner, 9 (14.5%) as a journal editor and 4 (6.4%) as a patient. The majority of respondents lived in North America (52.8%) and Europe (34.2%). When comparing survey responses by role, we focus here on those categorised as researchers, practitioners and policymakers, as these were well-represented decision-maker subgroups who serve unique knowledge user roles. Full characteristics of respondents are summarised in online supplemental appendix D, table 1.

The vast majority of respondents reported they were familiar with SRs (621/684 (90.8%)). Two survey questions were included to further gauge the respondents' understanding of SRs (online supplemental appendix 1, table 2 Q2-3). The first question asked the following: 'According to Moher and colleagues, the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) checklist is not a quality assessment instrument to judge the quality of an SR'. This question was ambiguous as PRISMA is a reporting checklist to determine the comprehensiveness of *reporting* of a published SR manuscript (and not the methodological quality) (Q2). However, the majority (335/542 (61.8%)) correctly agreed that PRISMA was not used to assess methodological quality (of conduct) of an SR. The second question asked respondents to identify limitations that occurred at the level of the SR (Q3). Slightly more than half (308/547 (56.3%)) correctly answered that all options except risk of bias of the primary studies applied. More than a third (199/547 (36.4%)) incorrectly answered that all but selective reporting of results and analyses applied. During peer review, we were asked to conduct a post hoc analysis to explore the potential for difference in responses according to the survey respondent type: 234 respondents from the targeted decision-makers' email list compared with the unknown respondents from social media. When stratifying by type of respondent for the PRISMA question (Q2), 150/214 (70.1%) of email recipients answered correctly, as opposed to 185/362 (51.1%) of participants that clicked the anonymous link on social media. For Q3, 125/214 (58.4%) of email recipients correctly answered the question about biases at the SR level compared with 183/362 (50.6%) of the anonymous link respondents.

Experiences and barriers to choosing SRs when more than one exists on the same topic

Respondents (n=558) often (64.5%) or sometimes (32.6%) sought out SRs as a source of evidence in their decision-making (never did=2.9%; online supplemental appendix D, table 3). Respondents (n=538) reported facing a situation where they found more than one SR on a given topic sometimes (54.8%) or often (43.1%) (table 1 Q6).

Just over half (50.8%) responded that they often or sometimes struggled to choose the most valid and trustworthy SR among multiple SRs on the same topic (table 1 Q9). Overall, the most common barrier to making this decision was reported to be a lack of time to fully read and evaluate each SR (55.2%) (table 1 Q10). Other frequent barriers were related to variation in the quality of conduct of SRs (54.2%), differences in results and conclusions across SRs (49.7%), variation in the primary studies included in the SRs (44.6%) and slightly different clinical focuses of the SRs (43.1%) (table 1 Q10). Of interest, when asked why they struggled to pick one SR, 35.6% (180/505) of respondents said it was because there was insufficient data from titles and abstracts to assess relevance to their question. Additionally, 27% of respondents recognised their inexperience assessing the methodological quality of SRs as being a barrier.

The reported approach to choosing the most appropriate SR tended to vary, depending on the type of decision-maker (table 1 Q8):

- Practitioners most often (75/170 (44.1%)) chose the most recently published SR(s) that were relevant to their topic. About one quarter (45/170 (26.4%)) found as many as they could that were relevant to their topic and then reviewed them all, and one tenth

Table 1 Considerations when there are multiple SRs on the same topic

Item	Responses	ALL	Policymaker	Practitioner	Researcher
Q6. How often have you faced a situation where you find more than one SR on a given topic of interest to you?	Never	(n=538)	(n=62)	(n=167)	(n=266)
		12 (2.2%)	0 (0%)	4 (2.3%)	5 (1.9%)
	Sometimes	295 (54.8%)	30 (48.3%)	106 (63.4%)	123 (46.2%)
	Often	232 (43.1%)	31 (50.0%)	57 (34.1%)	138 (51.9%)
Q8. When you encounter multiple SRs on the same topic how do you choose the one(s) most likely to address your clinical/public health/policy question or your learning needs?	I typically choose the first one I find that is relevant to my topic	(n=552)	(n=62)	(n=170)	(n=266)
		13 (2.4%)	1 (1.6%)	2 (1.1%)	4 (1.5%)
	I find as many as I can that are relevant to my topic and then review them all	207 (37.5%)	34 (54.8%)	45 (26.4%)	111 (41.7%)
	I typically choose the most recently published one(s) that are relevant to my topic	171 (31.0%)	7 (11.2%)	75 (44.1%)	60 (22.6%)
	I typically choose the one from the highest impact factor journal	34 (6.2%)	1 (1.6%)	18 (10.6%)	10 (3.8%)
Q9. When you have encountered multiple SRs on the same topic, which of the following statements resonates most with you?	I can usually identify the SR(s) best suited to my needs	(n=548)	(n=62)	(n=170)	(n=264)
		268 (48.9%)	35 (56.4%)	56 (32.9%)	152 (57.6%)
	I sometimes struggle to identify the SR(s) that are best suited to my needs	238 (43.4%)	24 (38.7%)	94 (55.2%)	101 (38.2%)
	I often struggle to identify the SR(s) best suited to my needs	41 (7.4%)	3 (4.8%)	19 (11.1%)	10 (3.8%)
Q10. If/when you struggle to choose the SR(s) best suited to your needs, the barriers to you being able to make this decision are	Insufficient data from titles and abstracts to assess relevance to my question	(n=505)	(n=60)	(n=165)	(n=251)
		180 (35.6%)	21 (35.0%)	55 (33.3%)	77 (30.7%)
	Inexperience with assessing the methodological quality of (or biases in) SRs	140 (27.7%)	9 (15%)	72 (43.6%)	30 (12.0%)
	Not enough time to read each SR in full to evaluate all the options	279 (55.2%)	23 (38.3%)	110 (66.7%)	119 (47.4%)
	You don't trust the conclusions	56 (11.0%)	11 (18.3%)	21 (12.7%)	34 (13.5%)
	Different results and conclusions across the SRs	251 (49.7%)	28 (46.7%)	94 (57.0%)	130 (51.8%)
	Variation in the quality of how the SRs were conducted	274 (54.2%)	34 (56.7%)	79 (47.9%)	146 (58.1%)
	Variation in searches across the SRs	172 (34.0%)	23 (38.3%)	46 (27.9%)	95 (37.8%)
	Variation in included primary studies across the SRs	225 (44.6%)	30 (50.0%)	63 (38.1%)	120 (47.8%)
	Variation in how across the SRs results were synthesised	194 (38.4%)	28 (46.7%)	51 (30.9%)	106 (42.2%)
	Slightly different clinical focus between SRs	218 (43.1%)	27 (45.0%)	74 (44.8%)	107 (42.6%)

*Numbers do not add up to 100% because respondents may have chosen more than one response option, and the majority of respondents identified as more than one type of decision maker (eg, researcher and patient).
SRs, systematic reviews.

(18/170 (10.6%)) chose the SR from the highest impact factor journal.

- Policymakers (34/62 (54.8%)) and researchers (111/266 (41.7%)) most often reported finding as many SRs as possible relevant to their topic of interest and reviewed them all.

The majority of decision-makers (335/385 (87.0%)) responded that they would use a free, automated, AI-informed, evidence-based online tool to assist in choosing the best SR(s) among multiple on the same question, if it was available (figure 2). Of respondent subgroups, practitioners reported the highest interest (90.5%), followed by researchers (86.9%), then policymakers (79.5%).

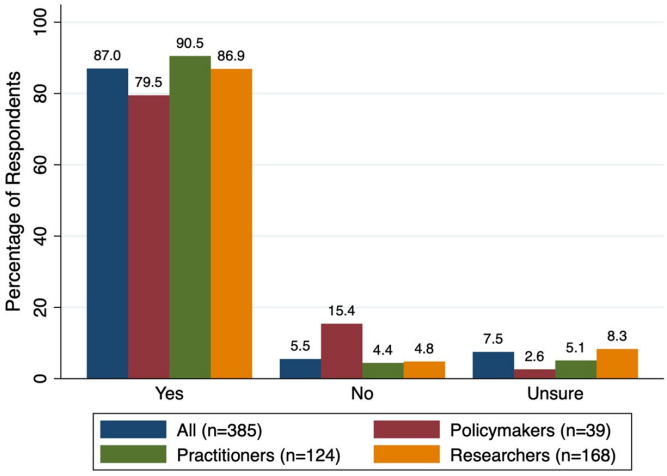

Figure 2 If a free, automated, tool was available to assist you in choosing the best systematic review(s) among multiple on the same question, would you use it? (Q7).

Data elements to consider when choosing the SRs best suited to my needs from multiple on the same topic

The two survey questions asking respondents to identify important SR features either via a prepopulated list (Q11) or determined through free-form response to a case study (Q12a–c) were randomised to 274 and 278 respondents and answered by 274 and 186 respondents, respectively. See online supplemental appendix D, table 4 for complete answers.

Q11—pre-defined features provided via drop-down menu

The relevance of the SR's research question to the respondent's clinical question or learning needs was selected most frequently as an important feature (83.6%), followed by the methodological quality and reproducibility of the SR (79.2%). The SR search strategy was also identified a key consideration, with recency of the SR search date (74.8%), and comprehensive search strategy (69.0%) was frequently selected. Other features more than half of respondents recognised as important to consider were: (1) the relevance of clinical outcomes (65.7%), (2) having a risk of bias assessment conducted for primary studies (60.9%), (3) a published protocol or preregistration for the SR (59.1%) and (4) consideration for the types of studies included (ie, randomised controlled trials vs non-randomised) (55.1%) (online supplemental appendix D, table 4 Q11).

Q12—free-form responses identifying features considered when choosing between SRs relevant to a case

The case study was based on the clinical question 'Is acupuncture effective/efficacious and safe for women with primary dysmenorrhea'. The PICO, characteristics and features of three SRs (Lui[40]; Yu *et al*[41]; Woo *et al*[42]) were presented in a table. Respondents chose the SR by Woo 2018 (86.6%) most frequently based on its strengths and weaknesses (online supplemental appendix D, figure 4 Q12a). Four out of 10 policymakers (44.8%) indicated that they would use all three in their decision-making.

When asked which criteria helped the respondents decide, the most common response involved using a hierarchy of features, not just one feature, to choose between SRs (or elements of SRs) (73.1%). The next most common consideration was the number of studies or patients included (45.2%). This response was most common among practitioners, about two-thirds, compared with one-third of researchers, and one-tenth of policymakers. The recency of the SR search date (29.6%), the risk of bias of the primary studies being assessed (28.5%) and the methodological quality of the SR (26.3%) were the next most common responses (online supplemental appendix D, table 4 Q12b).

Additional features identified by free-form responses and not included in the prepopulated elements (Q11) were: heterogeneity, and the process for selecting, extracting and assessing studies (online supplemental appendix D, table 4 Q12b).

DISCUSSION

We surveyed policymakers, practitioners, researchers, learners and other respondent types to understand how they use SRs to inform decision-making, and how they compare and select one or more SRs when there are multiple addressing the same clinical, public health or policy question. These individuals, who demonstrated good baseline knowledge of SRs, often sought out SRs as a source of evidence in their work and decision-making. They were also frequently faced with a situation where they find more than one SR on a given topic of interest. Nearly half of all respondents have struggled to choose the most valid and trustworthy SR, most often due to lack of time, and owing to varying methodological quality of identified SRs, variability in the primary studies included and differences in their results and conclusions. The proposed use of an AI tool to assist in comparing multiple SRs on the same topic was well accepted by survey respondents.

When comparing SRs on the same topic, themes of important characteristics were related to the relevance of the SR to their PICO question of interest, the robustness of the literature search and included studies, the recency, and the methodological quality. However, the answers varied by type of decision-maker, where healthcare practitioners more often chose the most recently published review(s) relevant to their topic, and policymakers and researchers most often reviewed all the relevant SRs on their topic of interest based on a hierarchy of criteria. This may indicate that practitioners are happy with a decision support tool to compare features of SRs, which presents the 'bottom line' synthesis or ranking of the SR evidence, but policymakers and researchers want the distilled information of all the presented reviews from which to make their own methodological judgments.

Another identified theme was that there is usually not one single best SR to ultimately choose. While a review may have a good AMSTAR-2 quality rating,[43] it still may contain important flaws like failing to report patient

important outcomes (eg, adverse events). It may also miss data and information relevant to the decision-maker. The three SRs in our case study contained different primary studies and publication dates, making their comparison especially difficult. To get the most out of the data for their decision-making, respondents (especially policymakers) commented that they often would review and read all the SRs on a topic and assess the strengths and weaknesses before making any decisions.

Strengths and limitations

A strength of our research was that we conducted it in accordance with an *a priori* published protocol. We combined newsletter, email distribution lists and social media to reach a wide range of decision-makers from across the globe. We attempted to maximise the response rate by sending email reminders and repeating messages through social media. Social media circulation as a distribution method proved fruitful although we had no control over exposure. This is the first study to our knowledge that explores decision-makers' interest in AI tools for navigating evidence from SRs.

A limitation was that we were expecting a 25% response rate to email invitations but achieved less than 10%. Another limitation is that we were unable to calculate a true survey response rate since two-thirds (n=450) responded through social media links, which were anonymous. Survey fatigue is a significant issue with this form of research and a multimodal approach to maximise reach, even at the expense of being able to calculate a response rate, was deemed necessary. Additionally, the piloting phase of our survey was conducted by eight academics. Had we included a more diverse group of decision-makers at this stage we may have improved the applicability of survey questions and overall response rate.

Response bias in our sample was also a major limitation as decision-makers working in higher income countries were over-represented. Additionally, the sample represented in this survey, namely individuals involved in guideline development and policymaking, may have been more likely to have responded. Another limitation is that our targeted emails and social media advertisement may have missed other important decision makers that use SRs.

Implication for practice, policy and knowledge translation

With the number of SRs on the same topic growing exponentially each year, we can predict that the challenge of decision-makers struggling to compare and choose between SR evidence will continue to increase. Our survey suggested that currently, over one-third of decision-makers reviewed only the data in the title and abstract when making their choice of SR (as opposed to the full text) when faced with clinical or policy decisions. Titles and abstracts may not contain enough information to make informed choices between SRs, and the full-text publications should be sought to enable methodological quality/risk of bias assessment and a full comparison of the strengths and weaknesses across SRs on the same topic.

It is important to note that an AI tool may be useful for comparing and deciding between SRs but cannot overcome methodological limitations of the SRs themselves. Researchers play an important role designing high-quality studies, as do journals in ensuring that the highest quality methods are used in research publications. This stresses the importance of capacity-building opportunities for researchers, peer reviewers and journal editors on conducting and evaluating SR manuscripts.

Future research

We have identified a role for an AI tool to help decision-makers more efficiently compare and choose SR evidence, and preferred SR features be the focus of the tool. Preferences for AI assistance appear to vary depending on the type of decision-maker, and we strive to develop a tool that is suitable to the needs of all users. The WISEST AI tool, in development, will present the scope, strengths and limitations of all the SRs found on the users' topic as well as a methodological quality ranking.[32] Users can then compare the different SRs and make their own clinical and methodological judgement about which SRs are most suited to their needs. The WISEST AI tool will aim to provide a 'supporting' role for decision-makers when comparing and choosing between reviews and not a substitution role.[32]

Our survey identified that there is a considerable proportion of SR users who feel they have inadequate assessment skills to adequately assess SR methodological quality. We also identified that practitioners, policymakers and researchers struggle to thoroughly understand biases at the SR level compared with the primary study level. The availability of the WISEST AI tool will help distill relevant SR information but is not intended to replace critical assessment or judgement. Appropriate supporting education, resources and training coupled with the user-friendly tool is necessary to overcome this knowledge gap.

Conclusions

Policymakers, practitioners and researchers often sought out SRs as a source of evidence in their decision-making, and often encountered more than one SR on a given topic of interest. Just over half of those surveyed struggled to choose the most valid and trustworthy SR among multiple. These struggles related to lack of time, and difficulty comparing different SRs when they vary in methodological quality and characteristics. When comparing SRs on the same topic, relevance to the question of interest, robustness and recency of the search, and methodological quality of the SR were most important to respondents. The development and implementation of an AI tool to rapidly highlight the features, strengths and weaknesses of SRs will help address these challenges and facilitate healthcare decision-making.

Author affiliations
[1]Knowledge Translation Program, Li Ka Shing Knowledge Institute, UBC, Toronto, Ontario, Canada
[2]Evidence Synthesis, Precisionheor LLC, Vancouver, British Columbia, Canada
[3]Faculty of Medicine and Health Sciences, McGill University, Montreal, Québec, Canada
[4]Department of Pharmacy, Nova Scotia Health Authority, Halifax, Nova Scotia, Canada
[5]Yealth Network, Beijing Health Technology Co., Ltd, Beijing, China
[6]Department of Biomedical and Neuromotor Sciences, University of Bologna, Bologna, Italy
[7]Anesthesiology, Pharmacology & Therapeutics, The University of British Columbia, Vancouver, British Columbia, Canada
[8]Institute for Health Services and Health System Research, Faculty of Health Sciences Brandenburg, Brandenburg Medical School Theodor Fontane, Neuruppin, Brandenburg, Germany
[9]Department of Pharmacy, Ottawa Hospital, Ottawa, Ontario, Canada
[10]Faculty of Medicine, University of Ottawa, Ottawa, Ontario, Canada
[11]Li Ka Shing Knowledge Institute of St Michael's Hospital, Knowledge Translation Program, St Michael's Hospital, Toronto, Ontario, Canada
[12]Institute of Health Policy Management and Evaluation, University of Toronto, Toronto, Ontario, Canada
[13]University of Ottawa, Ottawa, Ontario, Canada
[14]Knowledge Translation Program, Li Ka Shing Knowledge Institute, St Michael's Hospital, Toronto, Ontario, Canada
[15]Department of Family Practice, The University of British Columbia—Vancouver Campus, Vancouver, British Columbia, Canada
[16]Knowledge Translation Program, Li Ka Shing Knowledge Institute, University of Toronto, Toronto, Ontario, Canada
[17]Department of Electrical and Computer Engineering, Toronto Metropolitan University, Toronto, Ontario, Canada
[18]Knowledge Translation Program, St Michael's Hospital, Toronto, Ontario, Canada

X Carole Lunny @carole_lunny, Sera Whitelaw @serawhitelaw and Emma K Reid @emma_k_reid

Contributors CL conceived of the study design and wrote the study protocol. CL, SW, YC, JHZ, DP, SK, NF, BS, A-AV, CA, BP, EKR, EB and ACT revised the study design protocol. CL and SW entered the survey into Qualtrics and managed the responses. CL analysed the data. CL edited the manuscript. A-AV and JD created tables and figures. CL, SW, YC, JHZ, JD, DP, SK, NF, BS, A-AV, CA, BP, EKR, EB and ACT edited and approved of the final manuscript. CL acted as guarantor.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient and public involvement Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Patient consent for publication Not applicable.

Ethics approval This study involves human participants but University of British Columbia (ID H20-02013) exempted this study. Participants gave informed consent to participate in the study before taking part.

Provenance and peer review Not commissioned; externally peer-reviewed.

Data availability statement All data relevant to the study are included in the article or uploaded as supplementary information. All data are contained in the manuscript and appendices.

ORCID iDs
Carole Lunny http://orcid.org/0000-0002-7825-6765
Emma K Reid http://orcid.org/0000-0002-7980-1673
Yuan Chi http://orcid.org/0000-0001-7454-3970
Nicola Ferri http://orcid.org/0000-0002-8563-9967
Andrea C Tricco http://orcid.org/0000-0002-4114-8971

REFERENCES

1 Sackett DL. Evidence-based medicine. *Semin Perinatol* 1997;21:3–5.
2 Higgins JP. Cochrane handbook for systematic reviews of interventions. John Wiley & Sons, 2019.
3 Ioannidis JPA. The mass production of redundant, misleading, and Conflicted systematic reviews and Meta-Analyses. *Milbank Q* 2016;94:485–514.
4 Ioannidis JPA. Why most published research findings are false. *PLOS Med* 2005;2:e124.
5 Jo CL, Burchett H, Bastías M, *et al*. Using existing systematic reviews for developing vaccination recommendations: results of an international expert workshop. *Vaccine (Auckl)* 2021;39:3103–10.
6 Bornmann L, Mutz R. Growth rates of modern science: A Bibliometric analysis based on the number of publications and cited references. *Asso for Info Sci & Tech* 2015;66:2215–22.
7 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:e1000326.
8 Ioannidis JPA. The mass production of redundant, misleading, and Conflicted systematic reviews and meta-analyses. *Milbank Q* 2016;94:485–514.
9 Taito S, Kataoka Y, Ariie T, *et al*. Assessment of the publication trends of COVID-19 systematic reviews and randomized controlled trials. *ACE* 2021;3:56–8.
10 Hoffmann F, Allers K, Rombey T, *et al*. Nearly 80 systematic reviews were published each day: observational study on trends in epidemiology and reporting over the years 2000-2019. *J Clin Epidemiol* 2021;138:1–11.
11 Abbott R, Bethel A, Rogers M, *et al*. Characteristics, quality and volume of the first 5 months of the COVID-19 evidence synthesis Infodemic: a meta-research study. *BMJ Evid Based Med* 2022;27:169–77.
12 Dotto L, Kinalski M de A, Machado PS, *et al*. The mass production of systematic reviews about COVID-19: an analysis of PROSPERO records. *J Evid Based Med* 2021;14:56–64.
13 Nothacker J, Stadelmaier J, Siemens W, *et al*. Characteristics of registered and published systematic reviews focusing on the prevention of COVID-19: a meta-research study. *BMJ Open* 2022;12:e060255.
14 Page MJ, Moher D. Mass production of systematic reviews and Meta-Analyses: an exercise in Mega-Silliness. *Milbank Q* 2016;94:515–9.
15 Moher D. The problem of duplicate systematic reviews. *BMJ* 2013;347:bmj.f5040.
16 Lunny C, Neelakant T, Chen A, *et al*. "Bibliometric study of 'Overviews of systematic reviews' of health interventions: evaluation of prevalence, citation and Journal impact factor". *Res Synth Methods* 2022;13:109–20.
17 Bougioukas KI, Vounzoulaki E, Mantsiou CD, *et al*. Global mapping of Overviews of systematic reviews in Healthcare published between 2000 and 2020: a Bibliometric analysis. *J Clin Epidemiol* 2021;137:58–72.
18 Jadad AR, Cook DJ, Browman GP. A guide to interpreting discordant systematic reviews. *CMAJ* 1997;156:1411–6.
19 Li Q, Wang C, Huo Y, *et al*. Minimally invasive versus open surgery for acute Achilles tendon rupture: a systematic review of overlapping meta-analyses. *J Orthop Surg Res* 2016;11:65.
20 Mascarenhas R, Chalmers PN, Sayegh ET, *et al*. Is double-row rotator cuff repair clinically superior to single-row rotator cuff repair: a systematic review of overlapping meta-analyses. *Arthroscopy* 2014;30:1156–65.
21 Zhao JG, Wang J, Long L. Surgical versus conservative treatments for displaced Midshaft Clavicular fractures: A systematic review of overlapping meta-analyses. *Medicine (Baltimore)* 2015;94:e1057.
22 Lunny C, Thirugnanasampanthar SS, Kanji S, *et al*. How can Clinicians choose between conflicting and discordant systematic

reviews? A replication study of the Jadad algorithm. *BMC Med Res Methodol* 2022;22:276.
23 Harris RG, Neale EP, Ferreira I. When poorly conducted systematic reviews and meta-analyses can mislead: a critical appraisal and update of systematic reviews and meta-analyses examining the effects of Probiotics in the treatment of functional constipation in children. *Am J Clin Nutr* 2019;110:177–95.
24 Llanaj E, Muka T. Misleading meta-analyses during COVID-19 pandemic: examples of methodological biases in evidence synthesis. *J Clin Med* 2022;11:4084.
25 Lucenteforte E, Moja L, Pecoraro V, *et al*. Discordances originated by multiple meta-analyses on interventions for myocardial infarction: a systematic review. *J Clin Epidemiol* 2015;68:246–56.
26 Hill A, Mirchandani M, Ellis L, *et al*. Ivermectin for the prevention of COVID-19: addressing potential bias and medical fraud. *J Antimicrob Chemother* 2022;77:1413–6.
27 Lawrence JM, Meyerowitz-Katz G, Heathers JAJ, *et al*. The lesson of Ivermectin: meta-analyses based on summary data alone are inherently unreliable. *Nat Med* 2021;27:1853–4.
28 O'Mathúna DP. Ivermectin and the integrity of Healthcare evidence during COVID-19. *Front Public Health* 2022;10:788972.
29 Mhaskar R, Emmanuel P, Mishra S, *et al*. Critical appraisal skills are essential to informed decision-making. *Indian J Sex Transm Dis AIDS* 2009;30:112–9.
30 Petticrew M. Why certain systematic reviews reach uncertain conclusions. *BMJ* 2003;326:756–8.
31 Page MJ, McKenzie JE, Kirkham J, *et al*. Bias due to selective inclusion and reporting of outcomes and analyses in systematic reviews of randomised trials of Healthcare interventions. *Cochrane Database Syst Rev* 2014;2014:MR000035.
32 Lunny C, Thirugnanasampanthar SS, Kanji S, *et al*. Protocol and plan for the development of the automated algorithm for choosing the best systematic review, 2021. Available: https://osf.io/nbcta
33 Eysenbach G. Improving the quality of web surveys: the checklist for reporting results of Internet E-surveys (CHERRIES). *J Med Internet Res* 2004;6:e34.
34 Dillman DA. Mail and Internet Surveys: The Tailored Design Method--2007 Update with New Internet, Visual, and Mixed-Mode Guide. John Wiley & Sons, 2011.
35 Gupta K. A practical guide to needs assessment. John Wiley & Sons, 2011.
36 Dillman DA. *Mail and Internet Surveys*. 2nd edn. Hoboken, New Jersey: John Wiley & Sons Inc, 2007:18.
37 Lwanga SK, Lemeshow S. Sample Size Determination in Health Studies: A Practical Manual. World Health Organization, 1991. Available: https://iris.who.int/handle/10665/40062
38 Keating JL, McKenzie JE, O'Connor DA, *et al*. Providing services for acute low-back pain: a survey of Australian Physiotherapists. *Man Ther* 2016;22:145–52.
39 Walker BF, French SD, Page MJ, *et al*. Management of people with acute low-back pain: a survey of Australian Chiropractors. *Chiropr Man Therap* 2011;19:29.
40 Liu T. Acupuncture for primary Dysmenorrhea: A meta-analysis of randomized controlled trials. *Alt Ther Health Med* 2017;23.
41 Yu S-Y, Lv Z-T, Zhang Q, *et al*. Electroacupuncture is beneficial for primary Dysmenorrhea: the evidence from meta-analysis of randomized controlled trials. *Evid Based Complement Alternat Med* 2017;2017:1791258.
42 Woo HL, Ji HR, Pak YK, *et al*. The efficacy and safety of Acupuncture in women with primary Dysmenorrhea: a systematic review and meta-analysis. *Medicine (Balt)* 2018;97:e11007.
43 Shea BJ, Reeves BC, Wells G, *et al*. AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of Healthcare interventions, or both. *BMJ* 2017;358:j4008.
44 Canadian Institute for Health Research. Knowledge user engagement. Canadian Institute for Health Research, 2016.
45 Lunny C, Veroniki AA, Hutton B, *et al*. Knowledge user survey and Delphi process to inform development of a new risk of bias tool to assess systematic reviews with network meta-analysis (rob NMA tool). *BMJ Evid Based Med* 2023;28:58–67.
46 Boutron I, Ravaud P. Misrepresentation and distortion of research in BIOMEDICAL literature. *Proc Natl Acad Sci U S A* 2018;115:2613–9.
47 Whiting P, Savović J, Higgins JPT, *et al*. ROBIS: a new tool to assess risk of bias in systematic reviews was developed. *J Clin Epidemiol* 2016;69:225–34.
48 Page MJ, McKenzie JE, Bossuyt PM, *et al*. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *Syst Rev* 2021;10:89:89:.
49 EQUATOR Network. EQUATOR (Enhancing the QUAlity and Transparency of health Research) Network, 2022. Available: https://www.equator-network.org/about-us
50 Shea BJ, Grimshaw JM, Wells GA, *et al*. Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews. *BMC Med Res Methodol* 2007;7:10.
51 Lorenz RC, Matthias K, Pieper D, *et al*. A Psychometric study found AMSTAR 2 to be a valid and moderately reliable appraisal tool. *J Clin Epidemiol* 2019;114:133–40.
52 Pollock M, Fernandes RM, Hartling L. Evaluation of AMSTAR to assess the methodological quality of systematic reviews in Overviews of reviews of Healthcare interventions. *BMC Med Res Methodol* 2017;17:48.
53 Balshem H, Helfand M, Schünemann HJ, *et al*. GRADE guidelines: 3. rating the quality of evidence. *J Clin Epidemiol* 2011;64:401–6.