Bench Marks!

Software Tools for Literature Screening in Systematic Reviews in Biomedical Research

Stevie van der Mierden[1], Katya Tsaioun[2], André Bleich[1,#] and Cathalijn H. C. Leenaars[1,3,#]

[1]Institute for Laboratory Animal Science, Hannover Medical School, Hannover, Germany; [2]Evidence-based Toxicology Collaboration at Johns Hopkins Bloomberg School of Public Health (EBTC), Baltimore, MD, USA; [3]Faculty of Veterinary Sciences, Utrecht University, Utrecht, The Netherlands

Abstract

Systematic reviews (SRs) hold promise for implementing the 3Rs in animal sciences: they can retrieve available alternative models, help refine experiments, and identify insufficiencies in, or an excess of, scientific knowledge on a particular topic. Unfortunately, SRs can be labor- and time-intensive, especially the reference screening and data extraction phases. Fortunately, several software tools are available that make screening faster and easier. However, it is not always clear which features each tool offers. Therefore, a feature analysis was performed to compare different reference screening tools as objectively as possible. This analysis enables researchers to select the tool that is most appropriate for their needs.

Sixteen different tools were compared: CADIMA, Covidence, DistillerSR, Endnote, Endnote using Bramer's method, EPPI-Reviewer, EROS, HAWC, Microsoft Excel, Excel using VonVille's method, Microsoft Word, Rayyan, RevMan, SyRF, SysRev.com, and SWIFT Active Screener. Their support of 21 features categorized as mandatory, desirable, and optional was tested.

DistillerSR, EPPI-Reviewer, Covidence, and SWIFT Active Screener support all mandatory features. These tools are preferred for screening references, but none of them are free. The best scoring free tool is Rayyan, which lacks one mandatory function: distinct title/abstract and full-text phases. The lowest scoring tools were those not specifically designed for SRs, like Microsoft Word and Endnote. Their use can only be advised for small and simple SRs.

A well-informed selection of SR screening tools will benefit review quality and speed, which can contribute to the advancement of the 3Rs in animal studies.

Introduction

Worldwide, millions of animals are used every year for research. In 2016, over 1.2 million mice were used in the UK[1] and over 1.4 million in Germany[2] for animal experimentation. The 3Rs (Russell and Burch, 1959) aims to reduce the number of animals used, refine animal experiments to cause less pain, suffering and distress, replace animal experiments with non-animal methods, and to increase the quality of animal research. There are many different approaches that are being utilized to implement the 3Rs. One such approach is the systematic review (SR), which is the focus of this paper.

An SR is a protocol-driven literature review that addresses a specific research question by collecting all relevant papers on the topic and extracting and analyzing their data in a transparent and objective manner. The SR results in a qualitative data analysis and may result in a quantitative meta-analysis (de Vries et al., 2014). SRs are already standard practice in clinical sciences. They are considered to provide the highest level of evidence (Hooijmans et al., 2010) and serve as a foundation of evidence-based prac-

[1] https://www.gov.uk/government/statistics/statistics-of-scientific-procedures-on-living-animals-great-britain-2016

[2] https://www.bmel.de/DE/Tier/Tierschutz/_texte/TierschutzTierforschung.html?docId=10323474#doc10323474bodyText1

# contributed equally

Received February 13, 2019; Accepted May 7, 2019; Epub May 10, 2019; © The Authors, 2019.

ALTEX 36(3), 508-517. doi:10.14573/altex.1902131

Correspondence: Stevie van der Mierden, PhD, Institute for Laboratory Animal Science, MHH, Carl-Neuberg-Straße 1, 30625, Hannover, Germany (vandermierden.stevie@mh-hannover.de)

tice guidelines in medicine. The Cochrane[3] organization facilitates and ensures high quality SRs in clinical research and public health, which are achieving high societal impacts. In 2016, 90% of WHO guidelines were based on Cochrane SRs[4].

The application of SRs in animal sciences is still at an early stage, but the number of SRs published in animal sciences is increasing (van Luijk et al., 2014). SRs can assist animal sciences in several ways, such as helping to identify gaps in current scientific knowledge (Hooijmans et al., 2010) or showing where enough evidence is available and new animal experiments are no longer necessary, thus preventing redundant animal studies. For example, an SR plus cumulative meta-analysis by Sena et al. (2010a) showed that the effect of tissue plasminogen activator in animal models of stroke was known, but still new studies using animals were being performed.

SRs can also guide the refinement of animal experiments. First, they can be used to assess the reporting of the methodological quality of published papers, assessing the experimental design elements such as randomization and blinding. Papers that do not specify randomization or blinding may report inflated effect sizes (Macleod et al., 2008; van der Worp et al., 2007). Second, SRs can be used to determine the extent of publication bias. Publication bias may mean that experiments on animals have been performed, but the results are not available to the public as negative outcomes are often not published. Besides, publication bias can lead to overestimation of efficacies (Sena et al., 2010b). This may lead to unnecessary animal experiments and even to the initiation of human clinical trials based on incorrect assumptions. SRs are, therefore, becoming increasingly important in laboratory animal science.

Tools for performing systematic reviews

The important role that SRs play in clinical research and (evidence-based) medicine is in part due to the rigorously validated methodology. This methodology includes several distinct phases of performing an SR. The most time-consuming are 1) developing an extensive search strategy in multiple databases to collect as many potentially relevant papers as possible, 2) screening of the papers for relevance in two phases (title plus abstract and full text) by at least two independent reviewers, and 3) extracting all relevant data of the included papers. For an extensive description of the full process, please refer to the Cochrane handbook (Higgins and Green, 2011).

Depending on the research question, the number of papers that need to be screened can range from a few hundred (Hirst et al., 2013) to thousands (Wever et al., 2015). The time needed to screen a title and abstract can vary greatly, with reports ranging from twelve seconds median (Bramer et al., 2017), to 30 seconds average (Higgins and Deeks, 2008), to one minute average (Shemilt et al., 2016). Depending on the size and complexity of the SR and on the experience of the reviewers, the total screening time can run into hundreds of man-hours. Besides, the screening phase provides logistical challenges such as blinding the reviewers and keeping track of overall progress without losing track of papers. The screening phase has been described to be amongst the most difficult and most time-consuming parts of an SR, and the part most in need of a good support tool (Carver et al., 2013). Fortunately, many such tools are now available to systematic reviewers, which can make screening proceed faster, more easily, and more efficiently. However, it is not easy to determine the differences between tools.

Several articles have compared different SR tools before (e.g., Kohl et al., 2018; Marshall et al., 2014). However, these focus on the complete SR and provide little detail on the screening phase.

Because of the potential time investment and the great need for a tool during the screening phases, we compared different SR tools as objectively as possible. To this end, we used a feature analysis based on the DESMET method for software evaluation (Kitchenham et al., 1997). In this method, a list of features is established *a priori*. In our case, these are features that help to perform screening of papers in an SR. Each tool is individually assessed to determine which features they support. The result is a table that shows which tools support which features. Our table enables researchers interested in performing an SR to choose a suitable tool for fast, efficient, and reliable screening.

SR tools analyzed

To select the tools to be assessed, the authors intended to perform a scoping search in PubMed to review what tools recent systematic reviews in animal sciences had used. However, of the thirty scoped hits (of which 27 were SRs), only one review mentioned which tool had been used, while 48% of papers reported the software used for the statistical analysis. Instead, the authors used a combination of feedback from their expert network and the systematic review toolbox website[5] (Marshall), which is a web-based catalogue of tools made for performing systematic reviews, to identify the relevant tools. An additional tool (EPPI-Reviewer) came to our attention only when this article was in preprint, and was included at this stage because of its potentially high relevance.

With one exception, all suggested tools were analyzed: CADIMA (Kohl et al., 2018), Covidence[6] (Veritas Health Innovation), DistillerSR[7] (Evidence Partners), Endnote[8] (Clarivate

[3] https://www.cochrane.org

[4] https://www.cochrane.org/news/use-cochrane-reviews-inform-who-guidelines

[5] http://systematicreviewtools.com/index.php

[6] https://www.covidence.org/home

[7] https://www.evidencepartners.com/products/distillersr-systematic-review-software/

[8] https://endnote.com/

Analytics), Endnote using the method described by Bramer et al. (Bramer et al., 2017), EROS[9] (Early Review Organizing Software, Instituto de Efectividad Clínica y Sanitaria), HAWC[10], EPPI-Reviewer[11], Microsoft Excel[12] (Microsoft Corporation), Microsoft Excel using the method described by VonVille[13], Microsoft Word[14] (Microsoft Corporation), Rayyan (Ouzzani et al., 2016), RevMan 5[15] (The Cochrane Collaboration), SyRF[16] (CAMARADES-NC3Rs), SysRev[17] (Insilica), and SWIFT Active Screener[18] (Sciome). The R package "Metagear" was suggested, but not further analyzed due to repetitive installation issues.

Bramer described a method for using Endnote, and VonVille described one for using Excel. These are both not tools by themselves, but they mitigate some of the inherent shortcomings of these non-specific tools. As both Endnote and Excel are readily available and popular software, these methods were also analyzed.

Feature analysis

The authors based the feature analysis on the qualitative method described in the DESMET project, which developed and validated a method for evaluating software engineering methods and tools (Kitchenham et al., 1997). According to the DESMET method, three steps have to be taken before the actual analysis can take place: 1) determine which features are to be analyzed, 2) give each feature a level of importance, and 3) determine the level of conformance to which features have to comply.

The list of features was established based on feedback from within the authors' networks and on internal discussion. Considerations on features to include comprised all features that are necessary to perform a systematic review according to the Cochrane guidelines (Higgins and Deeks, 2008). The authors also looked at the PRISMA statement (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) for possible features. The PRISMA statement mentions two possible features that are relevant for screening: allowing in-/exclusions with reasons for exclusion, and creating a flow diagram (Liberati et al., 2009). The features analyzed in this study are presented in Table 1.

For this analysis, the authors decided on three levels of importance for the features: mandatory, desirable, and optional. The authors consider the features categorized as mandatory to be the minimum a tool has to support in order to facilitate a high-quality systematic review that complies with the Cochrane guidelines. The features "multiple user support", "in-/excluding references", "distinct TiAb/full-text phases", and "discrepancy resolving" are directly coupled to chapter 7 of the Cochrane handbook "Selecting studies and collecting data" (Higgins and Deeks, 2008). The other mandatory features were selected to ensure the correct technical working of the tools, i.e., the tool is accurate, you can import and export all your data, and the available support makes it easy to use the tool correctly. A tool that does not completely support all mandatory features is not necessarily incapable of producing high-quality systematic reviews, but additional measures have to be implemented when performing the systematic review to ensure its quality. For example, with tools that do not support distinct in-/exclusion phases, it usually is possible to export the results of the title & abstract screening phase and import the included studies again for full-text screening. However, it is relatively easy to lose references or data when one exports and reimports data, especially when one manually copies and pastes the data.

The desirable and optional features are not strictly needed to create an SR in line with the Cochrane guidelines, but they help to perform an SR faster, more easily, and/or they increase the quality of the SR. Desirable features in general have a larger impact than optional features. The feature "Free to use" is notable, as it is not inherently a feature that a tool supports. However, since the cost can greatly influence the choice of a tool, it was deemed important enough to assign it to the desirable level.

The levels of importance decided for the different features are presented in Table 1. Nine mandatory, nine desirable, and three optional features were analyzed.

There are two types of features: dichotomous features, which consist of yes/no questions, and compound features, which consist of multiple possible levels of support for a feature. The minimal level of conformance to which the features must comply were established by discussion between the authors. Where appropriate, they were coupled to the COCHRANE guidelines. Meeting a conformance level means that the feature meets the minimum requirement to contribute to a high-quality SR. The authors thoroughly considered what the different features really entail, and what it would mean for these features to be supported. The possible levels of conformance for each feature are presented in Table 1.

For each feature of each tool, it first was established whether a certain feature was present or not. If a feature was present, its functioning was compared to the conformance levels described in Table 1. For example, for the feature "exporting results", it was first checked whether exporting of the results was possible with the tool under consideration. If exporting was possible, it was then checked into which file extensions exporting was possible.

For transparency reasons it is important to note that one of the levels of conformance was changed after the first analysis. The

9 http://www.eros-systematic-review.org/
10 https://hawcproject.org/
11 http://eppi.ioe.ac.uk/cms/
12 https://products.office.com/en-us/excel
13 https://shwca.se/Excel-SR-workbooks-guides
14 https://products.office.com/en-us/word
15 https://community.cochrane.org/help/tools-and-software/revman-5
16 http://syrf.org.uk/
17 https://sysrev.com
18 https://www.sciome.com/swift-activescreener/

Tab. 1: Features and levels of conformance
The table describes the features, their relative importance for the process they support (mandatory, desirable, optional), and whether a feature is dichotomous (simple) or compound. The numbers in brackets for compound features indicate the number of levels of conformance for that feature. The different levels at which a tool can support a feature and the minimum threshold for a feature to be considered supported are also given.

Feature	Importance of feature	Simple (S) or compound (C) feature	Description of levels of conformance	Acceptance threshold
Status of software	Mandatory	C (3)	**No longer supported:** The tool is no longer in development and/or errors are no longer patched. **Beta software:** The tool is actively developed but does not have a stable release. **Stable release:** The tool is mature, actively supported, and has a stable release.	Stable release
Customer support	Mandatory	C (3)	**No support:** Help documentation is inadequate (does not help to solve many questions/problems) and the company does not reply in a reasonable amount of time or does not help solve the issue. **Documentation only:** There is adequate documentation available but the company does not reply in a reasonable amount of time or does not help solve the issue. **Direct support:** There is adequate documentation available and the company replies in a timely manner and actively supports the customer by answering questions and helping with issues.	Documentation only
Multiple user support	Mandatory	C (3)	**No multiple user support:** It is not possible for multiple users to work at the same time, on the same project, independently from each other, and blinded. **Two user support:** Two users can work at the same time, on the same project, independently from each other, and blinded. **Multiple user support:** An unlimited number of users can work at the same time, on the same project, independently from each other, and blinded.	Two-user support
Reference importing	Mandatory	C (3)	**No formal import:** The tool does not formally support importing of references; references have to be entered manually (this includes copy-pasting). **Limited files supported or difficult process:** The tool can only import using a limited number of file extensions (e.g., only CSV) and/or the process is difficult. **Fully supported:** The tool has an easy process for importing references and supports multiple file extensions.	Fully supported
Reference allocation	Mandatory	C (4)	**No formal allocation:** There is no formal method for allocating references to reviewers. **Allocation possible:** It is possible to allocate references to reviewers, but the tool does not support randomization of this step. **Allocation + re-allocation:** The tool is able to re-allocate references to different reviewers (e.g.,s when a reviewer drops out).	Allocation + re-allocation
In-/excluding references	Mandatory	C (3)	**No system for in-/exclusion:** The tool has no formal system for in- or excluding references. **In-/exclusion only:** The tool supports in- and excluding references, but no reason for exclusion can be given. **In-/exclusion + reason for exclusion:** The tool supports in- or excluding of references, and a reason for exclusion can be given.	In-/exclusion + reason for exclusion
Distinct TiAb/ full-text phases	Mandatory	C (3)	**No distinct phases:** There is no clear distinction between the title/abstract phase and the full-text phase; there is only one phase. **TiAb & full-text phase:** There is a clear distinction between the title/abstract phase and the full-text phase. **User-defined phases:** The user can create as many distinct phases as they need.	TiAb & full-text phase
Discrepancy resolving	Mandatory	S	**No:** There is no official process to resolve discrepancies. **Yes:** Official support for discrepancy resolving.	Yes

Feature	Importance of feature	Simple (S) or compound (C) feature	Description of levels of conformance	Acceptance threshold
Exporting results	Mandatory	C (3)	**No export:** No formal export is supported, exporting must be done manually. **Limited export:** Support for formal export, but only in limited file extensions (e.g., only .txt or .xlsx). **Full export:** It is possible to export the results in at least the .CSV format, or multiple general file extensions are supported.	Limited export
Free to use	Desirable	S	**No:** The tool must be purchased or free/trial accounts have severe limitations that can compromise the systematic review, e.g., a strict time limitation (<1 year, only one user per project, limit on number of references accepted). **Yes:** The tool can be used for free and without practical limitations that can compromise the review.	Yes
Randomizing order of references	Desirable	S	**No:** It is not possible to randomize the order of references for the reviewers. **Yes:** It is possible to randomize the order of references for the reviewers.	Yes
Keyword highlighting	Desirable	C (3)	**No highlighting:** No keyword highlighting possible or highlighting of only one word is possible. **3rd party only:** The tool does not support formal keyword highlighting, but it is possible to use (free) 3rd party software for highlighting (e.g., extensions for Google Chrome, Add-ons for Firefox). **Highlighting possible:** The tool natively supports the highlighting of more than one word.	Highlighting possible
Multiple user roles	Desirable	C (4)	**No different roles:** There are no different roles for different users; everybody has the same role and rights in the project. **Reviewer + Manager roles:** Two different roles with different rights for reviewers and for manager roles. **Any further role:** The tool supports both reviewer and manager roles, but also any further roles (e.g., librarian role). **User definable roles:** The users can determine the number of roles and determine the rights for the roles.	Reviewer + manager
Project auditing	Desirable	S	**No:** The tool does not support auditing the project; a complete overview of all alterations by all users on the project. **Yes:** The tool supports auditing the project.	Yes
Non-Latin character support	Desirable	S	**No:** The tool does not support non-Latin characters (e.g., Cyrillic, Greek, Chinese, Arabic, etc.). **Yes:** The tool supports non-Latin characters.	Yes
Show project progress	Desirable	C (3)	**No project progress:** There is no way to determine the overall progress of the project (e.g., % completed) **Limited progress:** The tool only shows rudimentary project progress (e.g., only the total % of references completed/ still to do) **Detailed progress:** The tool can display detailed progress (e.g., the progress per reviewer)	Detailed progress
Attaching comments	Desirable	S	**No:** It is not possible to attach comments to references. **Yes:** It is possible to attach comments to references.	Yes
Attaching PDFs	Desirable	S	**No:** The tool does not support uploading PDFs for full text screening. **Yes:** The tool supports uploading PDFs for full text screening.	Yes
Reference labelling	Optional	S	**No:** It is not possible to attach *a priori* determined labels to references. **Yes:** It is possible to attach *a priori* determined labels to references.	Yes
Flow diagram creation	Optional	S	**No:** The tool cannot automatically provide a flow diagram meeting the PRISMA criteria. **Yes:** The tool can automatically provide a flow diagram meeting the PRISMA criteria.	Yes
Machine learning/ automation	Optional	S	**No:** The tool has no form of machine learning or automation of the screening process. **Yes:** The tool has a form of machine learning or automation of the screening process.	Yes

level of conformance for the feature "reference allocation" was initially defined as follows: "allocation + randomization: The tool can randomly allocate references to reviewers". We later felt that this might be too similar to the feature of "randomizing order of references". Therefore, we decided to focus rather on the option to have reviewers drop out, i.e., to be able to replace reviewers who can no longer participate in the screening phase of the SR.

One of the initial optional features was whether the tool was user-friendly. However, during the analysis the authors found no objective method to assess whether a tool is user-friendly or not. In the end, all tools were successfully used to test the screening of references, and thus all tools were considered user-friendly. Due to the subjective nature of this feature and the little information it provides, the authors decided to leave out this feature from the results.

Analysis of tools and data

The primary feature analysis was performed by one author (SvdM). All results were assessed by a second SR expert (KT or CHCL). Tools were accessed between July 27 and the September 3, 2018, except for the VonVille method for Excel, the Bramer method for Endnote, and EPPI-Reviewer, which were analyzed in April and May of 2019. All data were collected and analyzed using Microsoft Excel 2010.

Number of features supported

The feature analysis gives three numbers for how many features a tool supports: mandatory, desirable, and optional. An overview of the results is presented in Figure 1. Four of the tested tools, SWIFT Active Screener, Covidence, DistillerSR, and EPPI-Reviewer, met all mandatory (9) criteria. The number of mandatory features supported by the other tools ranged from eight to two. None of the tools reached the maximum number (9) of desirable features supported, with DistillerSR supporting the most (8) desirable features and EROS the fewest (2). Only DistillerSR and EPPI-Reviewer supported the maximum number (3) of optional features. Five tools supported two of the optional features: CADIMA, SWIFT Active Screener, Covidence, Rayyan, and HAWC.

Table 2 shows the features that were supported by the tools at a more granular level. Features with a numerical value in the cell are compound features for which there are different levels of conformance, as described in Table 1. A numerical value in the cells in Table 2 indicates the difference between the minimum conformance level of a feature (0) and the level of conformance that a particular tool supports. For example, for the "multiple user support" feature, the minimum level of conformance is "two user support". If a tool supports a maximum of two blinded users working independently, then the supported level of conformance is equal to the minimum level of conformance, and thus the number in Table 2 would be "0". Similarly, if the tool supports multiple blinded independent users working at the same time, the supported level of conformance would be one level higher than the minimum level, and the number in Table 2 would be "+1".

The customer support feature was supported by all tools, meaning that each tool has extensive help documentation, video tutorials, or access to a help desk. Two of the mandatory features were supported by almost all tools. Only EROS did not have a stable release (i.e., mature supported software). Exporting the results was supported by all tools except for SyRF and RevMan. The support for desirable features was more varied. Only two tools, SWIFT Active Screener and EROS, did not support non-Latin

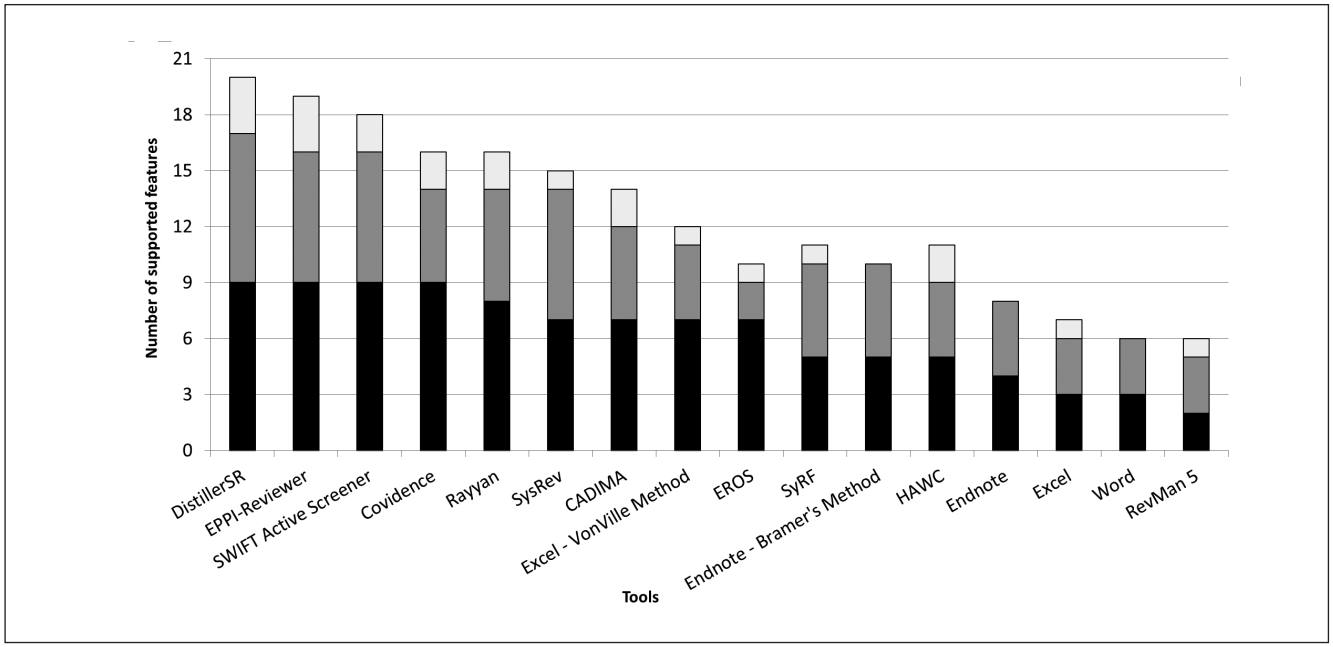

Fig. 1: The number of features supported by different tools
The number of mandatory features is presented in black, desirable features in light grey, and optional features in white.

Tab. 2: Overview of features supported by the software tools
Shaded cells indicate that the tool fully supports that specific feature. For compound features (features with more than one possible level of conformance), the number in the cell indicates the number of levels of conformance above or below the minimum level a feature is supported; 0 means that the feature matches the minimum level of conformance. For example, for the feature "distinct TiAb/full-text phases", a 0 means the tool supports distinct title/abstract and full-text screening phases, a -1 means that there are no separate phases, a +1 means that the user can perform more than two phases. The complete list of conformance levels is provided in Table 1.

	Mandatory									Desirable									Optional		
Tool	Status of software	Customer support	Multiple user support	Reference importing	Reference allocation	In-/excluding references	Distinct TiAb/Full-text phases	Discrepancy resolving	Exporting results	Free to use	Randomizing order of references	Keyword highlighting	Multiple user roles	Project auditing	Non-Latin character support	Show project progress	Attaching comments	Attaching PDFs	Reference labelling	Flow diagram creation	Machine learning /automation
CADIMA	0	+1	+1	-1	-1	0	0		0			-1	0			0					
Covidence	0	+1	+1	0	0	0	0		+1			0	-1			0					
DistillerSR	0	+1	+1	0	0	0	+1		+1			0	+2			0					
Endnote	0	+1	-1	0	-1	-2	-1		+1			-2	-1			-2					
Endnote – Bramer Method	0	0	0	0	-1	-1	-1		+1			-2	0			-2					
EPPI-Reviewer	0	+1	+1	0	+1	0	+1		+1			0	0			0					
EROS	-1	+1	+1	-1	0	0	0		0			-1	+1			-1					
HAWC	0	+1	-1	0	-1	0	-1		0			-1	+1								
Microsoft Excel	0	+1	-1	-1	-2	-2	-1		+1			-2	-1			-1					
Excel – Vonville method	0	0	0	-1	-1	0	0		+1			-2	0			-2					
Microsoft Word	0	+1	-1	-2	-2	-2	-1		0			-2	-1			-2					
Rayyan	0	+1	+1	0	0	0	-1		+1			0	+1			0					
RevMan	0	0	-1	-1	-2	-1	-1		-1			-2	-1			-2					
SyRF	0	+1	+1	-1	-1	0	-1		-1			-1	0			-2					
SysRev.com	0	+1	+1	-1	0	0	-1		0			-1	0			0					
SWIFT Active Screener	0	+1	+1	0	0	+1	+1		-1			0	0			0					

characters and only three tools, Rayyan, CADIMA, and HAWC, did not allow attaching comments to references. The support of optional features was varied. Only Microsoft Word, Endnote, and Endnote using Bramer's method did not allow the labelling of references. In contrast, only five tools supported the creation of flow schemes and four tools had some form of automation or machine learning.

Ranking of SR tools

The results of the feature analysis make it possible to rank the SR tools for screening based on the analysis used in this paper. In ranking, mandatory features supersede both desirable and optional features, and desirable features supersede optional features. The ranking of the analyzed tools is presented in Figure 1.

Of the four tools supporting all mandatory features, DistillerSR supports the most desirable features, followed by EPPI-Reviewer and SWIFT Active Screener, and then by Covidence. All four of these tools require a paid license. Of the free to use tools, Rayyan ranks the highest; the one mandatory feature it lacks is the support for distinct TiAb & Full-text screening phases. SysRev, CADIMA, VonVille's method for Excel, and EROS follow Rayyan in mandatory feature support, but SysRev also supports seven desirable features, compared to five for CADIMA, four for VonVille's method, and two for EROS.

Feature analysis for choosing software tools

To the authors' knowledge, this is the first feature analysis specifically for software tools supporting the screening phase of SRs. Feature analysis is a powerful method to compare different tools, as it allows for a systematic approach with as little bias as possible. We minimized bias by selecting the features to be tested and the criteria for assessing them before starting the analysis. We based our selection criteria on the Cochrane handbook (Higgins and Green, 2011) as far as possible. In addition, primary screening results were reviewed by at least one other independent reviewer. The result is a robust and objective evaluation of different tools, which enables an informed selection of the tool best suited to a project's specific needs.

Although feature analysis is a powerful tool, it does have some limitations. To start, most software is constantly under development, and the tested SR tools are no exception. This means that the number of supported features changes over time. For example, the developers of Rayyan and SyRF have stated that distinct title/abstract and full-text phases will be supported in future versions. The feature analysis was performed on the currently released version. It is unlikely, although possible, that future versions of the software will support fewer features; therefore, they are expected to include more features and score higher in future. It might, therefore, be worthwhile to contact the developers to inquire after recent developments, especially if the choice of a tool depends largely on a specific feature.

Other applications for feature analysis

The screening phase is one of the most difficult and time-consuming parts of the SR, and most in need of support tools. A good tool for screening thus has the highest impact on facilitating SRs. However, screening is not the only part of the SR that can benefit from tools. The data extraction phase is on par with the screening phase concerning required time and difficulty, and can also benefit greatly from tool support (Carver et al., 2013). A feature analysis could be helpful to identify the most appropriate tool for the data extraction phase.

Moreover, a tool that supports the entire SR process (question formulation, protocol development, automatic searches in multiple databases with import of the search results for screening of title and abstract and full-text, text-mining-enabled data extraction, meta-analysis, publication-ready PRISMA charts and tables generation) should be the ultimate goal for developers of SR tools. These features could greatly increase the number of SRs completed, which would ultimately make great contributions to science in all areas, including animal research and evidence-based transition to non-animal methods.

Highest scoring tools based on the scored features

Although there are, unfortunately, no tools that support all 21 tested features, we found four tools that support all the mandatory features: DistillerSR, EPPI-Reviewer, SWIFT Active Screener, and Covidence. Any of these tools can be used to ensure a high-quality screening phase of the SR, although they might lack features that make the process a bit faster or help keep an overview of the project's progress.

The four best scoring tools are not free to use. For DistillerSR, "free to use" is the only desirable feature not supported, making DistillerSR the best scoring tool overall according to the criteria defined in this analysis. EPPI-Reviewer does not support auditing. This means that if a project has multiple "manager" roles, it is not possible to verify afterwards who uploaded which references or who made alterations to the project. SWIFT Active Screener's main deficiency at this time is lack of support for non-Latin characters (e.g., Cyrillic characters). With English being the predominant language in science, this might not be the biggest deficiency, although it can be more problematic with older papers, which are more likely to be in the author's language in countries where many papers are published in a non-English language (e.g., China), and in certain fields where the use of mathematical formulae or words containing Greek letters (e.g., β-antagonists) is common. Of the four best scoring tools, Covidence, supports the fewest desirable features. Of particular note is the lack of support of the randomization of the order of references. This could potentially lead to bias of the reviewers in screening, such as fatigue bias or learning bias. Covidence also does not support multiple user roles. It is important to be able to allocate specific reviewers to such roles as the ability to import references, reports, set inclusion or exclusion criteria, or invite and remove reviewers. Lastly, Covidence has no auditing support.

The best scoring free tool is Rayyan. The only mandatory feature not supported by this tool is distinct title/abstract, and full-text phases. This can be circumvented by exporting all results after the title/abstract phase and importing them for the full-text phase. This can be cumbersome and, depending on the team, can lead to delays. For example, it is possible to have a person specifically tasked with full text screening, or data extraction, or risk of bias determination. In tools that support both phases, this person could start to work on this phase before the preceding phase has been completed and could work in parallel. When using tools that do not support separate phases, each phase would have to be completed before the next phase could commence.

The differences in levels of conformance do not affect the ranking of the highest scoring tools, as in our analysis the number of supported features outweighs the difference in levels of conformance, and there are no tools with the same number of supported features.

Lower scoring tools

The low scores of the lowest scoring tools can be explained by them not having been specifically developed to perform SRs. RevMan is designed to create a high-quality manuscript for an SR, but does not specifically help to perform the early steps. It lacks the features needed for screening independently and in a blinded manner. Microsoft Word and Microsoft Excel are general purpose software for word processing and spreadsheets, respectively. EndNote is a reference manager not designed for per-

forming the inclusion or exclusion of references or for multiple users working independently and in a blinded fashion.

As stated, Endnote and Excel are widely used in scientific research, but they are not specifically designed for SRs. In our analysis we included Bramer's method for Endnote and VonVille's method for Excel. Both of these methods adapt the respective software for SR reference screening. The main advantage of Bramer's method for Endnote is that it enables independent screening by multiple reviewers. VonVille's method substantially improves the workflow in Excel by enabling independent screening, in-/excluding of references, distinct screening phases, and discrepancy resolution. However, these methods cannot solve all the shortcomings of Excel and Endnote, and both of these methods still require manual copying and pasting of the references, which is prone to errors.

Choosing the appropriate tool

Which tool to use depends on three criteria: 1) available funding, 2) scope and/or difficulty of the specific SR, and 3) how many SRs are planned. If sufficient funding is available, DistillerSR, EPPI-Reviewer, SWIFT Active Screener, and Covidence are appropriate choices. If the SR is very small and straightforward (dozens of references instead of hundreds), EndNote using Bramer's method or Microsoft Excel using VonVille's method can be used. However, screening using these methods still has shortcomings, and if the SR is large or complicated these alternatives are too error-prone or cumbersome to recommend. For large or complicated SRs without funding, Rayyan is currently considered the best free option according to the analysis in this paper; its only major drawback is that it does not support distinct title/abstract and full-text phases, which may cause some delays, as explained above. The second-best free scoring tool SysRev is a good alternative, but currently it has a few major drawbacks: the first is one is, similar to Rayyan, it lacks distinct title/abstract and full-text phases. The second is that SysRev only supports one type of import file suitable for mass importing references (.xml). Finally, it does not include negative and positive keyword highlighting, which slows down the screening and increases user fatigue.

The SR is a powerful tool for implementing the 3Rs in animal science, but performing one requires substantial resources. Fortunately, tools are available to help perform systematic reviews. This paper helps scientists planning to perform an SR to make an informed decision on which tool is the most appropriate for their research needs. Our hope is that such informed decisions will result in the production of larger numbers of higher-quality systematic reviews, which can benefit the reduction, refinement, and replacement of animal studies.

References

Bramer, W. M., Milic, J. and Mast, F. (2017). Reviewing retrieved references for inclusion in systematic reviews using endnote. *J Med Libr Assoc 105*, 84-87. doi:10.5195/jmla.2017.111

Carver, J. C., Hassler, E., Hernandes, E. and Kraft, N. A. (2013). Identifying barriers to the systematic literature review process. *2013 ACM / IEEE International Symposium on Empirical Software Engineering and Measurement,* 203-212. doi:10.1109/esem.2013.28

de Vries, R. B., Wever, K. E., Avey, M. T. et al. (2014). The usefulness of systematic reviews of animal experiments for the design of preclinical and clinical studies. *ILAR J 55*, 427-437. doi:10.1093/ilar/ilu043

Higgins, J. and Deeks, J. J. (2008). Selecting studies and collecting data. In J. Higgins and S. Green (eds), *Cochrane Handbook for Systematic Reviews of Interventions: Cochrane Book Series* (151-185, Chapter 7). doi:10.1002/9780470712184.ch7

Higgins, J. and Green, S. (2011). *Cochrane Handbook for Systematic Reviews of Interventions*. https://www.handbook.cochrane.org

Hirst, T., Vesterinen, H., Sena, E. et al. (2013). Systematic review and meta-analysis of temozolomide in animal models of glioma: Was clinical efficacy predicted? *Br J Cancer 108*, 64-71. doi:10.1038/bjc.2012.504

Hooijmans, C. R., Leenaars, M. and Ritskes-Hoitinga, M. (2010). A gold standard publication checklist to improve the quality of animal studies, to fully integrate the three Rs, and to make systematic reviews more feasible. *Altern Lab Anim 38*, 167-182. doi:10.1177/026119291003800208

Kitchenham, B., Linkman, S. and Law, D. (1997). DESMET: A methodology for evaluating software engineering methods and tools. *Comput Control Eng J 8*, 120-126. doi:10.1049/cce:19970304

Kohl, C., McIntosh, E. J., Unger, S. et al. (2018). Online tools supporting the conduct and reporting of systematic reviews and systematic maps: A case study on cadima and review of existing tools. *Environmental Evidence 7*, 8. doi:10.1186/s13750-018-0115-5

Liberati, A., Altman, D. G., Tetzlaff, J. et al. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: Explanation and elaboration. *PLoS Med 6*, e1000100. doi:10.1371/journal.pmed.1000100

Macleod, M. R., Van Der Worp, H. B., Sena, E. S. et al. (2008). Evidence for the efficacy of NXY-059 in experimental focal cerebral ischaemia is confounded by study quality. *Stroke 39*, 2824-2829. doi:10.1161/strokeaha.108.515957

Marshall, C., Brereton, P. and Kitchenham, B. (2014). Tools to support systematic reviews in software engineering: A feature analysis. Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering. *ACM*, Article No. 13. doi:10.1145/2601248.2601270

Ouzzani, M., Hammady, H., Fedorowicz, Z. and Elmagarmid, A. (2016). Rayyan – A web and mobile app for systematic reviews. *Syst Rev 5*, 210. doi:10.1186/s13643-016-0384-4

Russell, W. M. S. and Burch, R. L. (1959). *The Principles of Humane Experimental Technique*. London, UK: Methuen.

Sena, E. S., Briscoe, C. L., Howells, D. W. et al. (2010a). Factors affecting the apparent efficacy and safety of tissue plasminogen activator in thrombotic occlusion models of stroke: Systematic review and meta-analysis. *J Cereb Blood Flow Metab 30*, 1905-1913. doi:10.1038/jcbfm.2010.116

Sena, E. S., van der Worp, H. B., Bath, P. M. et al. (2010b). Publication bias in reports of animal stroke studies leads to major overstatement of efficacy. *PLoS Biol 8*, e1000344. doi:10.1371/journal.pbio.1000344

Shemilt, I., Khan, N., Park, S. and Thomas, J. (2016). Use of cost-effectiveness analysis to compare the efficiency of study identification methods in systematic reviews. *Syst Rev 5*, 140. doi:10.1186/s13643-016-0315-4

van der Worp, H. B., Sena, E. S., Donnan, G. A. et al. (2007). Hypothermia in animal models of acute ischaemic stroke: A systematic review and meta-analysis. *Brain 130*, 3063-3074. doi:10.1093/brain/awm083

van Luijk, J., Bakker, B., Rovers, M. M. et al. (2014). Systematic reviews of animal studies; missing link in translational research? *PLoS One 9*, e89981. doi:10.1371/journal.pone.0089981

Wever, K. E., Hooijmans, C. R., Riksen, N. P. et al. (2015). Determinants of the efficacy of cardiac ischemic preconditioning: A systematic review and meta-analysis of animal studies. *PloS One 10*, e0142021. doi:10.1371/journal.pone.0142021

Conflict of interest

The authors have no conflict of interest to declare.

Acknowledgements

The authors would like to thank Julia M. L. Menon, Kim E. Wever, and Florenza L. Ripoli for their excellent feedback on which features and which tools to analyze. We also thank J. Thomas for his suggestion to add EPPI-Reviewer to our analysis. We acknowledge support by the German Research Foundation (DFG) and the Open Access Publication Fund of Hannover Medical School (MHH).