A Comparative Analysis of Support Techniques for Assessing the Quality of Systematic Literature Reviews

Rand Alchokr$^{1[0000-0003-0112-5430]}$, Athul Sunilkumar[1], Gunter Saake$^{1[0000-0001-9576-8474]}$, Thomas Leich$^{2[0000-0001-9580-7728]}$, and Jacob Krüger$^{3[0000-0002-0283-248X]}$

[1] Otto-von-Guericke University, Magdeburg, Germany
{rand.alchokr,Sunilkumar,saake}@ovgu.de
[2] Harz University & METOP GmbH, Wernigerode, Germany
tleich@hs-harz.de
[3] Eindhoven University of Technology, Eindhoven, The Netherlands
j.kruger@tue.nl

Abstract. The rapidly growing number of scientific publications poses numerous challenges for researchers engaged in literature analyses. Structured methodologies like systematic literature reviews are becoming increasingly expensive, considering their attempt to cover all relevant publications. Despite the increasing efforts needed, the importance of literature reviews also leads to an increasing growth in their number. While there are support techniques (e.g., guidelines, tools, checklists) for conducting literature analyses, a concise and clear overview of such techniques for assessing the quality of the analysis itself is missing. Such an overview can help researchers identify techniques for their work, understand ambiguities between them, support peer reviews, and guide future research by highlighting open gaps. In this paper, we address this lack of an overview by identifying existing techniques for assessing the quality of systematic literature reviews, comparing their properties, and discussing their pros and cons. For this purpose, we elicited 14 techniques through a systematic literature search covering 15 years (2007–2021). Overall, our contributions can help researchers identify feasible techniques for assessing the quality of literature analyses and can guide the development of new techniques, thereby facilitating the conduct and improving the quality of literature analyses.

Keywords: Systematic Literature Reviews · Quality Assessment · Computer Science · Software Engineering.

1 Introduction

The increasing amount of scientific publications is one of the most complex challenges for researchers. Several factors, such as resource constraints, competition, and publication bias, contribute to increasing publications and complexity. An overwhelming amount of publications also leads to expensive and time-consuming processes during literature analyses. Researchers perform such analyses to identify the most relevant studies related to a research topic, understand the consequent knowledge, and stay up-to-date. A Systematic Literature Review (SLR) is a dedicated form of literature analysis that follows a systematically structured methodology to analyze, aggregate, and ideally summarize all findings on a particular topic [28,44]. Stemming from the medical domain, SLRs have become one of the most popular research methodologies in computer science and other domains to elicit existing evidence. Considering computer science and particularly software engineering, Kitchenham [27,28] was among the first to explore the idea of using SLRs as a means towards Evidence-Based Software Engineering (EBSE). EBSE envisions researchers collecting the current "best" evidence on a research topic and integrating it with practical experiences as well as human values into decision-making frameworks for developing and maintaining software. This idea underpins the value of SLRs in providing a systematic overview of a research topic that can guide other researchers and practitioners [12,41].

Unfortunately, conducting an SLR is resource-intensive, since it involves complex, multi-faceted, and time-consuming processes that necessitates substantial dedication and technical expertise to ensure reliability. Many researchers are developing semi-automatic techniques to facilitate different stages of a SLRs [3,9,13,15,20,32,33,43,45,46,56]. However, despite such advances, we are unaware of analyses or automation that ensure the quality of reporting (systematic) literature reviews. Due to the importance of SLRs as a research methodology for collecting evidence, there has been a surge in the number of publications reporting them in all scientific domains [4,18]. Even though this is a good sign for a move towards more evidence-based research, the quality of these reviews has been a controversial matter. Besides potential technical problems [29,48], serious pitfalls arise considering aspects like defining, reporting, and justifying the search strategy, search procedure, source selection, quality assessment criteria, or result synthesis [52].

Furthermore, the constant increase in publications has raised concerns regarding the reliability of SLRs in terms of comprehensively covering the literature, given the original intent of ensuring "thoroughness" and "completeness". Because SLRs build on a set of guidelines for framing research questions, sorting out studies, and analyzing the quality of primary studies, their credibility depends on the strength of such guidelines—and most reviews still make use of outdated guidelines or tools. For instance, the Database of Abstracts of Reviews of Effects (DARE) criteria are often used, even though they have severe limitations regarding completeness [4]. Additionally, the reporting quality of SLRs varies, limiting their readers' ability to assess their strengths and weaknesses [34]. To improve the quality of SLRs (particularly in computer science), it is important to constantly monitor existing guidelines, identify their limitations, and refine them if needed to improve credibility.

To address these problems, we employed an automated search on Scopus to identify techniques for SLRs in different domains, leading to a total of 14 publications in a period of 15 years (2007–2021). We explore these publications and extract from each one: the type of quality assurance (e.g., guideline, tool, checklist) it contributes (RQ_1), the scientific domain it stems from (RQ_2), as well as the pros and cons associated with the proposal (RQ_3). We argue that researchers, particularly those in computer science, software engineering, and digital libraries, benefit significantly from designing support infrastructures for a unified quality-assessment framework for SLRs.

2 Background and Related Work

Ideally, every scientific endeavor begins with a literature review, through which researchers can gain a comprehensive understanding of previous work in their area of interest. Thus, they obtain a solid foundation while avoiding unnecessary repetitions. There are established research methods, such as SLRs or systematic mapping studies, to consolidate the existing knowledge or evidence regarding a specific problem. Such methods emphasize a critical reflection of existing knowledge and may be used to identify open gaps. However, a literature review has limited scientific value if not conducted properly. For this reason, SLRs have arguably become the most established and widely used type of literature review in any scientific field (e.g., medicine, software engineering) [4,18,28,61], favored due to the systematic and replicable method of collecting literature. Over time, different scientific fields have proposed individual adaptations of the methodology, trying to improve its conduct while also accounting for specifics of their field (e.g., collecting techniques in software engineering instead of medical studies in medicine) [28,41,42,58].

An SLR comprises three main phases, each divided into sub-steps [28]: Planning, Conducting, and Reporting. Unfortunately, any systematic literature analysis is labor-intensive, error-prone, and time-consuming, which encouraged researchers to develop numerous techniques and tools to provide support for (semi-)automating parts of the process [6,9,13,16,17,20,32,33,43,45,46,56]. Some examples for such tools are SLuRp [6], StArt [14,17], or SLR-Tool [16]. Interestingly, while many other steps of the process are researched, assessing the quality of an SLR itself (in contrast to quality assessing the primary studies) is a controversial and less researched matter. In fact, there is a lack of consensus on how authors and reviewers of a literature review should assess a review's quality [11]. This is likely connected to the many steps involved, asking authors and reviewers to have a detailed understanding of searching, selecting, reading, comparing, classifying, and assessing publications. Moreover, a literature review must be judged based on its comprehensiveness, the depths of its analysis, and how broadly the existing literature is covered [5]. Since SLRs are intended to build on a standardized, replicable, and unbiased method following a stringent protocol, their associated risks are often unnoticed. For instance, some tertiary studies emphasize the importance of knowing potential threats to the initial search and its reporting in SLRs to ensure reliable evidence, but have found that important details in published literature reviews are missing [7,25,29]. Additionally, research questions should be established before conducting the review, but this may not always be possible. Specifically, plenty of literature reviews claim that the review itself serves as a foundation for formulating meaningful research questions. In such cases, only after obtaining a detailed understanding, the researchers can identify shortcomings in the current research, which enables the formulation of more relevant and meaningful research questions [5]. Also, training researchers to conduct SLRs remains a challenge, since the lack of guidelines and a consequent agreement on what constitutes high quality can lead to very different results. Lastly, many SLRs seem to build on outdated guidelines that lack critical updates [30], leading to potential quality issues in the review and its reporting. Reflecting on all challenges associated with SLRs, it is questionable to justify that these provide the best quality outcome [5,24]. An overview and synthesis of existing techniques for quality-assuring SLRs can help mitigate such problems.

3 Methodology

In this section, we describe the design of our SLR. To conduct our SLR, we followed the guidelines by Kitchenham et al. [28], which are the most established guidelines in software engineering [29].

3.1 Planning

Goal and Need. We aimed to identify publications that contribute techniques specifically designed for assessing the quality of SLRs. These contributions include tools, guidelines, checklists, and other evaluative techniques aimed at ensuring rigor and reliability in review processes.
Research Questions. To achieve our goal, we defined the following three Research Questions (RQs):

RQ$_1$ *What techniques for assessing the quality of a SLR exist?*
Our objective for this research question is to retrieve primary studies related to existing techniques for assessing the quality of an SLR.

RQ$_2$ *What fields of research do these techniques stem from?*
For context, it is important to identify the field a technique stems from.

RQ$_3$ *What are the pros and cons of the existing techniques?*
Finally, we aim to provide a detailed understanding of the individual techniques we identified, comparing their properties, pros, and cons.

Search Strategy. To identify relevant publications, we employed an automated search on Scopus. We chose Scopus because it is considered a high-quality database and covers various publishers and research fields, for instance, ACM, IEEE, Springer, and Elsevier. Scopus indexes only peer-reviewed publications, which is why we can assume a certain level of quality for the retrieved publications. Additionally, compared to other databases, Scopus provides search features that support researchers conducting an SLR, such as specific filtering and a query-building mechanism [47].

Search String. We constructed our search string by collecting keywords closely related to our research questions.

```
“(‘approach’ OR ‘measurement’ OR ‘appraisal’ OR ‘checklist’ OR
‘reporting’) AND (‘instrument*’ OR ‘tool*’ OR ‘guideline*’ )
AND (‘systematic literature review*’ OR ‘systematic review*’ OR
‘SLR’) AND (‘quality’ OR ‘evaluation’ OR ‘assessment’)”
```

We argue that this search string is feasible to identify publications that are relevant to our research questions. Note that it is specific to Scopus and would require adaptations for other search engines. After we employed our automated keyword search, we followed up with a backwards snowballing [58, 59] to complement and extend our search strategy—avoiding typical technical problems of automated searches and following established recommendations [29, 48].

Selection Criteria. To answer our research questions, we defined Inclusion Criteria (IC) and Exclusion Criteria (EC) for filtering relevant primary studies. We defined the following four inclusion criteria:

IC$_1$ The publication is reviewed and published in a scientific journal, conference, or workshop.
IC$_2$ The publication is available in `pdf` format.
IC$_3$ The publication has been published between 2007 and 2021.
IC$_4$ The publication is concerned with quality assessing SLRs.

Furthermore, we defined four exclusion criteria:

EC$_1$ The publication is not written in English.
EC$_2$ The publication is a presentation or abstract only.
EC$_3$ The publication is a thesis, technical report, or similar work.
EC$_4$ The publication lacks publisher information or publication type.

Quality Assessment. We assessed the quality of each primary study to rate its importance and relevance for our research questions and to properly synthesize our findings. In detail, we answered the following six questions (Q) using the scoring of 1 for “Yes,” 0.5 for “Partial,” and 0 for “No” based on the recommendations by Kitchenham [23]:

Q$_1$ *Is there a clear statement about the goal of the reported research?*
Q$_2$ *Are the technique's design decisions justified?*
Q$_3$ *Is the procedure of how the research has been conducted thoroughly explained?*
Q$_4$ *Are the evaluation results thoroughly analyzed and reported?*
Q$_5$ *Does the evidence support the findings presented?*
Q$_6$ *Is the method used to obtain the results feasible?*

We remark that we employed these quality criteria on different types of publications (e.g., guidelines, tools). So, our ratings are subjective and can also vary depending on the type of research.

Data Extraction and Analysis. We extracted the publications from Scopus as a `csv` file and imported that file into a spreadsheet. To select publications,

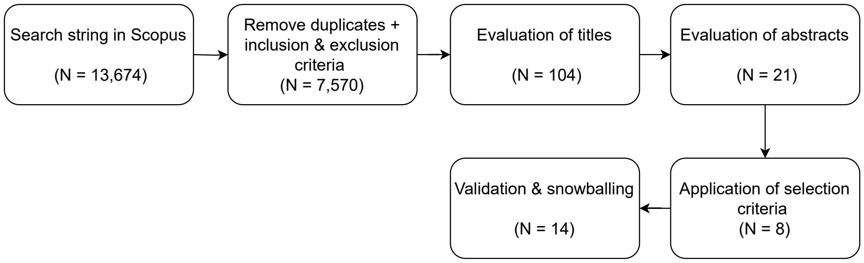

Fig. 1: Our process for identifying primary studies.

we first read each title and then the abstract to exclude those that do not fulfill our selection criteria. If we did not make an assessment, we read the respective publication completely. In case of any conflicts, we rechecked the corresponding publications and, in a few cases, we had to decide by majority vote. During snowballing, we added the newly identified publications to the spreadsheet. Afterwards, the second author extracted the relevant data (for consistency), while the others reviewed the results. This spreadsheet contains each publication's bibliographic data, the domain, name, and type (i.e., tool, guideline) of the technique, the phases that are targeted by the proposed method, and the pros and cons of each technique.

3.2 Conduct

We executed our literature search following the protocol we described in Section 3.1. We illustrate the steps of our conduct in Figure 1. First, we executed our search string on Scopus, retrieving 13,674 publications as a `csv` file that we stored as a shared spreadsheet. Then, the first two authors inspected each publication independently to identify those relevant to answering our research questions based on the selection criteria we defined. Please note that if we had any doubts about a publication in Phase 2 or Phase 3, we included it rather than excluding it, as recommended by Okoli and Schabram [38]. We ended up with 14 publications. From the snowballing step, we identified six of these primary studies, which are P9–P14. Afterwards, the second author extracted all relevant data into the shared spreadsheet, which the first author fully validated to check for correctness and completeness. Finally, we performed a cross-validation of the selected publications, extracted data, and re-iterated the snowballing (including the newly found publications), which did not lead to new publications.

4 Results

Next, we present the results for each of our research questions.

Table 1: Summary of the retrieved publications on assessing the quality of SLRs.

Id	Ref	Year	Name	Domain	Specific Field	Target	Type	Access	Function
P1	[49]	2007	AMSTAR	Medical	Healthcare Randomized Trials	Plan, Conduct, Report	Checklist	Open	Manual
P2	[30]	2009	PRISMA	Medical	Healthcare Randomized Trials	Report	Checklist	Open	Manual
P3	[36]	2013	BPMN	General	General	Plan, Conduct, Report	Tool	Restricted	Automated
P4	[24]	2013	Kitchenham QA Reporting	Computer Science	Software Engineering	Report	Checklist	Open	Manual
P5	[57]	2014	Eco Evidence	General	Environmental Studies	Plan, Conduct, Report	Tool	Restricted	Semi-automated
P6	[5]	2015	Being Systematic in SLRs	General	General	Plan, Conduct, Report	Guideline	Open	Manual
P7	[51]	2016	Mixed Filter	Medical	Mixed Health Care Studies	Plan, Conduct, Report	Guideline	Open	Manual
P8	[50]	2017	AMSTAR-2	Medical	Health Randomized & Non-Randomized Trials	Plan, Conduct, Report	Checklist	Open	Manual
P9	[22]	2017	DARE by CDR	General	General	Report	Checklist	Open	Manual
P10	[19]	2018	MMAT	General	Mixed Studies	Plan, Conduct, Report	Checklist	Open	Manual
P11	[2]	2018	CASP (2018)	Medical	Healthcare Randomized Trials	Report	Checklist	Restricted	Manual
P12	[4]	2019	CATSER	Computer Science	Software Engineering	Report	Tool	Open	Manual
P13	[40]	2020	PRISMA (2020)	Medical	Healthcare Randomized Trials	Report	Checklist	Open	Manual
P14	[31]	2020	CASP (2020)	Medical	Healthcare Randomized Trials	Report	Checklist	Restricted	Manual

4.1 RQ$_1$: Existing Techniques

To address RQ$_1$, we examined each selected publication in detail. Through this process, we were able to gather valuable insights and summarize the key properties of each technique—which we specify in Table 1. Precisely, we provide an overview of all techniques, including the publication's reference, year of publication, name of the technique, the domain from which it stems, its specific field, its main assessment target, its type, its accessibility, and the degree of automation. In the following, we briefly introduce each of the techniques.

P1 – AMSTAR [49]. The *Assessment of Multiple SysTemAtic Reviews* was developed in 2007 based on formerly introduced studies to assess the quality and to appraise SLRs critically. More specifically, this technique concerns healthcare research, including SLRs of randomized trials for which the conducting researchers must assess the trials' reliability and validity. An expert group was formed using the nominal group technique to validate the checklist and finalize its features. Initially, AMSTAR included 37 items presented as a checklist or questionnaire. Through the nominal group technique, these items were reduced to 29, which measured 11 components. While these 11 components cover the entirety of an SLR, they only assess how it is reported. In particular, the formulated research questions, the search for relevant studies, the assessment of primary-study quality, the data extraction, and the reporting of the findings are checked.

P2 – PRISMA [30]. The *Preferred Reporting Items for Systematic reviews and Meta-Analyses* have the primary objective of improving the reporting of SLRs.

Specifically, the goal is to define transparency standards, but not to propose quality criteria for the actual conduct. PRISMA is mainly useful when reviewing health-care interventions and helps readers to assess the trustworthiness and applicability of an SLR. This technique intends to improve the technique *QUality Of Reporting Of Meta-analyses* (QUOROM) [35], which is argued to possess poor reporting quality guidelines and to focus on meta-analyses of randomized controlled trials. In contrast, PRISMA was introduced to encompass both SLRs and meta-analyses. The technique follows Population Intervention Control Group Outcome and Study Design (PICOS) and encompasses a 27-item checklist and a 4-phase flow diagram.

P3 – BPMN [36]. The *Business Process Modeling Notation* is one of the few techniques that aims to fully automate a part of an SLR. So, its main objective is to reduce the time needed to conduct an SLR, contributing to the productivity and quality of conducting such reviews. Modeling the workflow of an SLR with the BPMN consists of three phases that map to the stages of the SLR itself: planning, conducting, and reporting.

P4 – Kitchenham Quality Assessment Reporting [24]. This technique aims to address quality issues within published studies that followed the guideline by Kitchenham and Charters [26]. For this purpose, 68 publications between 2005 and 2012 were identified, pertaining to the field of software engineering. The proposed 12-item quality assessment checklist was based on the guideline used [26]. A weighted scoring mechanism was used to derive a final quality checklist, resulting in an updated version of the guidelines [28]. These guidelines represent a standard quality assessment technique in software engineering SLRs.

P5 – Eco Evidence [57]. *Eco Evidence* aims to aid SLRs related to environmental science. This technique provides a structured standard report to ensure transparency and reproducibility. It is ideal for evaluating cause-effect relationships across diverse study types and for synthesizing results from both observational and experimental research. Eco Evidence combines a structured causal criteria approach, a weighted scoring system for evidence quality, and a database to store the published studies. It has an analyzer desktop tool that synthesizes the studies to test cause-effect hypotheses.

P6 – Being Systematic in SLRs [5]. This is a set of guidelines that mainly addresses the cons of conducting SLRs. Since many other publications focus on and support mitigating cons in this context, we concentrate on the claims in these guidelines that connect to the quality of an SLR. Particularly, it claims that objectivity and replicability are the main properties for assessing the quality of an SLR, which must be justified based on the transparency of the process.

P7 – Mixed Filter [51]. The *Mixed Filter* is intended to work for mixed empirical studies with different designs. It was applied to six journals from three fields: primary care, medical informatics, and public health/epidemiology. These fields were picked because they require addressing complex research questions with various research methods and designs of empirical studies. The performance of the Mixed Filter was analyzed using descriptive statistics drawn from the measurement of overall precision, sensitivity, and specificity across the six journals, which were the total relevant records that were drawn.

P8 – AMSTAR-2 [50]. This is an updated version of AMSTAR that was developed in 2017, almost 10 years after the introduction of its predecessor. It aims to extend the scope of AMSTAR to non-randomized trials in the healthcare domain. AMSTAR-2 includes 16 critical items. The main differences of this extension are the orientation of the research questions along the PICOS framework and improvements regarding the clarity of inclusion and exclusion criteria.

P9 – DARE by CDR [22]. The *Database of Abstracts of Reviews of Effects by the Centre for Reviews and Dissemination* provides a checklist that aims to evaluate the quality of SLRs related to multiple domains such as Software Process Improvement. Within the checklist, the planning and conduct phases are only considered based on their reporting. Overall, the checklist is based on four main questions with scores to assess the credibility, reliability, and quality of an SLR. Unlike more detailed techniques, DARE is accessible, easy to apply, and focused on methodological transparency [1].

P10 – MMAT [19]. The *Mixed Methods Appraisal Tool* was developed to appraise SLRs related to mixed studies, qualitatively and quantitatively. This tool considers five core methodological quality assessment design criteria: 1) Qualitative, 2) Randomized controlled trial, 3) Non-randomized, 4) Quantitative descriptive, and 5) Mixed methods studies. It includes five criteria for each of the above-mentioned main criteria. These are rated as "yes," "no," or "can't tell". The checklist of this tool helps a user fix the judging criteria. A table is provided for each category of study design that presents a detailed definition, designs, approaches, and explanation of the specific criterion. MMAI was developed based on already existing techniques, which were simplified.

P11 & P14 – CASP (2018 / 2020) [2, 31]. The *Critical Appraisal Skills Programme* tool has two versions developed in 2018 and 2020, respectively. It is designed to be used as an educational or research-related tool. The first version was released as a requirement for conducting workshops, but without a scoring system. Moreover, it was designed exclusively for randomized trials. A group of reviewers was assigned to develop and control the items in this checklist and to associate these with a workshop format. For each of the 10 checklist questions, a response is classified as "yes," "no," or "can't tell." The resulting list with 10 items was found to be meaningful in several contexts [2]. Still, the main focus of CASP is data synthesis concerning techniques proposed in previous research [54], which involves three stages: line-by-line coding, development of descriptive themes, and generation of analytical themes.

P12 – CATSER [4]. The *Critical Appraisal Tool for Software Engineering Systematic Reviews* is based on AMSTAR-2 and is still under testing. CASTER aims to raise awareness regarding the criticality of quality assessing SLRs in software engineering. This is a community-collaborated method towards the development of a tool, which led to the insight that existing tools for such quality assessments mainly expand upon tools used in the healthcare domain. However, CASTER also raises the point that almost no effort has been made to improve these tools to keep pace with advances in the healthcare industry either.

P13 – PRISMA (2020) [40]. Considering the advancements in science and technology, and consequently in SLR methodologies, an update within the PRISMA guidelines was deemed necessary. After reviewing the methods employed in 220 publications and conducting a survey among 21 member teams, the new PRISMA (2020) was proposed. So, the updated PRISMA extensively relies on community experiences and feedback from co-authors of over 15 SLRs. There are several enhancements in this version, such as an updated checklist, justification of any possible alternative data-synthesis methodologies used, citation of all excluded studies, or addressing the source of data, code, and other materials used in SLRs. Finally, a revision of the flow diagram template to adapt it to the updated guidelines has been proposed.

On $\mathbf{RQ_1}$. We identified 11 different techniques that have been proposed for assessing the quality of SLRs in the last 15 years. Please note that AMSTAR, CASP, and PRISMA exist in two versions each. All of the techniques are tools, guidelines, checklists, or a mixture of these, and are often applicable to a variety of literature analyses. The techniques mostly cover the reporting of an SLRs, with a few considering the conduct and reporting for assessing the quality. Positively, most techniques are publicly available.

4.2 $\mathbf{RQ_2}$: Fields of Research

SLRs in different domains (i.e., broader research areas) have, logically, varying sets of process steps. We provide an overview of all domains in Table 1. To identify these domains, we elicited which ones or which fields are explicitly mentioned in the respective publications. In some cases, we could clearly derive this information from the publication's context. For instance, we assigned the specific field of software engineering for the guideline by Kitchenham and Charters [26], which is within the broader domain of computer science. The same applies to the other techniques, for example, we assigned AMSTAR [49] to the specific field of healthcare randomized trials within the broader medical domain.

On $\mathbf{RQ_2}$. We can see that most techniques (7) relate to the medical domain, specifically healthcare randomized trials. The remaining seven publications cover four different domains, namely computer science (3), mixed studies (1), environmental studies (1), and all domains (2). For example, DARE by CDR is used mainly in healthcare and public policy research, but is adaptable to other domains like software engineering, education, and social sciences. The medical domain is more advanced regarding the use of SLRs, and its community is more accustomed to utilizing respective quality assurance techniques.

4.3 $\mathbf{RQ_3}$: Pros and Cons

To answer this research question, we examined each publication, paying close attention to the weak points of each technique as well as its advantages. We searched whether the researchers explicitly mentioned this information. As some

Table 2: Summary of the pros and cons of the retrieved techniques.

Id	Name	Pros	Cons
P1	AMSTAR [49]	• Ensures strong content validity through expert input and existing validated tools.. • Exploratory factor analysis refines and strengthens the tool by identifying essential items [39]. This ensures the extendability to non-randomized trials (healthcare).	• Requires more studies to confirm reproducibility and construct validity. • Quantifying bias remains difficult due to variability across study types and domains.
P2	PRISMA [30]	• Promotes clarity, transparency, and structured reporting in SLRs. • Adaptable to randomized and non-randomized studies. • Clearly outlines review protocols, information sources, and search strategies to support reproducibility. • Provides detailed insights into bias risks and helps readers in replication and updates of reviews.	• Is not recommended as a quality assessment tool, because it does not elaborate on the methodologies used in the SLR process. • The checklist itself was not developed through an SLR. • Evidence exists for some of the checklist items. • Inconsistent study quality, with some trials lacking clear reporting, and Small trials may be overestimated.
P3	BPMN [36]	• Based on an enhanced version of concrete guidelines proposed by [26] that contribute to quality. • It is automated, so it helps in reducing the time required to perform an overall SLR process. • It serves as a base for the development of further computational tools in the future.	• Its documentation cannot be considered top-notch, since it does not provide much information about practical scenarios. • Uncertainty about how the tool can adjust to the complexity of the studies. • It does not discuss important criteria like risks of bias.
P4	Kitchenhan QA reporting [24]	• Addresses many major problems concerning the SLR reporting process. • Gives improvement scope for the future of quality evaluation of software engineering studies based on empirical methods.	• Poor agreement on the study content quality as the scoring mechanism was found to be error-prone, which makes the validation process somewhat stringent. • The use of an extractor checker for data extraction from broad lessons and surveys increases the risk of missing important issues or misinterpretation of the same.
P5	Eco Evidence [57]	• Offers a causal criteria-based summary and detailed reporting and uses a weighted rating mechanism to assess study strength [37]. • Applicable to theoretical and practical research contexts. • Features a reusable evidence bank in an open-access database, reducing workload and simplifying evidence extraction.	• Restricted to environmental studies. • Manual evidence scoring can be time-intensive. • Requires domain expertise. • It contributes a minimum regarding a quality assessment that aids the reporting phase.
P6	Being Systematic in SLR [5]	• Discusses in detail the key attributes that should be considered to enhance the SLR process and proposes a change in existing guidelines. • Considers a very wide variety of study domains. • Analytical, and consideration of the impacts of bias.	• Covers the entire SLR process and does not focus on the reporting phase.
P7	Mixed Filter [51]	• Gives high performance even though it varies by journal. • The filter exhibits a high sensitivity factor, which means it could retrieve almost all relevant studies. • Shows a high specificity and precision of over 60 %.	• Tested across limited journals, and thus could result in varied results across other journals.
P8	AMSTAR-2 [50]	• Extends AMSTAR for non-randomized trials. • Provides improved bias assessment and highly justified study design criteria. • Enhances clarity through defined critical domains. • With cautious use, it can support teaching and act as a checklist for SLRs. • Enables deeper performance quality assessment to identify flaws in poorly conducted reviews.	• Lacks explanation of systematic review methods; relies on the Cochrane Handbook for full guidance [21]. • Improper handling of bias may lead to inaccurate impact estimation. • There is no proper specification of risk of bias tools for non-randomized trials, leaving it to reviewers' discretion.
P9	DARE by CDR [22]	• Helps to minimize bias and ensure maximization of the overall validity of the review. • Provides a needed breakthrough in assessing SLRs related to Software Process Improvement. • Functions based on concrete guidelines provided by DARE checklist [1]. • Provides a scope of improvements and the possibility of extending the tool over other domains as well.	• Since the literature is weak, the testing of the technique could also come with limitations. With growing literature, better testing could be carried out to improve the current version. • High risks of bias exist that researchers could frequently extract wrong data. • Risk of missing some related keywords in search strings or over-constraints in the same could result in loss of publications.
P10	MMAT [19]	• Serves as a concrete technique for Mixed Studies Reviews (MSR) as it provides methodological quality assessment criteria across various study designs. • It focuses on a limited number of core criteria, and it is very time-efficient.	• Focuses more on methodological quality rather than the reporting quality; thus, it is difficult to judge the criteria needed. • It is difficult to ensure the trustworthiness of methodological quality. • A revision of this technique is necessary to ensure content validity and reliability.
P11,14	CASP (2018), (2020) [2, 31]	• A large community assigned for ongoing evaluation and refinement. • Both versions (original and updated) remain relevant and improved, with added questions and responses [31]. • Higher-quality studies contribute to the technique. • The possibility of misinterpretation of the "tick box" checklist was addressed.	• The inferences of the technique were drawn based on a single case study, wherein it was made to appraise studies on semi-structured interview methodologies. • When covering various study types, the technique would yield different results. • Even though it is advised to use CASP over mixed datasets, it is still believed to be less feasible.

Id	Name	Pros	Cons
P12	CATSER [4]	• Developed collaboratively to tailor AMSTAR-2 for software engineering SLRs, retaining its strengths and all the advantages. • Addresses bias risk of SLR, reporting quality, and assessment using the latest guidelines available.	• A prototype that is under construction, and thus requires testing, refinements, and validation. • If not properly planned, all the cons of AMSTAR-2 could apply, too.
P13	PRISMA (2020) [40]	• Properly documented updates and refinement through collective reviewer feedback. • Includes an expanded checklist to reflect advances in SLRs. • Clearly describe and justify alternative data synthesis methods, and cite excluded studies with reasons.	• Limitations in the ability to collect feedback to a small extent. • Reporting guidelines addressing the presentation and synthesis of qualitative data should also be consulted [55]. • Possible risks in bias assessment.

techniques have been extended, updated, or published in different versions (e.g., PRISMA, CASP), we investigated the most recent version and list its pros and cons in Table 2. Our objective was to elicit the techniques used in prominent SLR quality assessment techniques, and, finally, to note their strengths and weaknesses. In Table 2, we summarize our findings.

On RQ$_3$. The updated versions of AMSTAR, PRISMA, Kitchenham Guidelines, and CASP outperform their former versions. Thus, we consider the older versions to be infeasible for contemporary SLRs. One common weakness of most techniques is bias detection and handling. As this is considered a complicated issue that involves multiple social aspects, numerous studies are devoted to investigating this direction as an important quality assurance aspect. Furthermore, comparing the techniques adopted in the medical domain, we can observe that the updated versions of AMSTAR and PRISMA support a wider variety of healthcare trials than CASP. In contrast, CASP is best for assessing individual studies, not reviews. AMSTAR, Kitchenham QA, and Eco Evidence focus on quality and evidence strength, while PRISMA and DARE focus on reporting and inclusion criteria. The most promising and reliable techniques appear to be the updated versions of AMSTAR and PRISMA. Both stand out due to their wider applicability and adaptability and have been improved to support a broader range of healthcare trials, making them versatile for medical research. Moving to computer science, a promising prototype, CATSER, adopted the advantages of the latest version of AMSTAR and utilized it to benefit software engineering in the same sense.

5 Discussion

In Table 2, we list the advantages and disadvantages of each technique. As we can see, the disadvantages differ widely between individual techniques. However, we can conclude that none of them is perfect and we would consider none optimal for all use cases. For example, in the software engineering domain, the first set of guidelines for SLRs was proposed by Kitchenham and Charters [26]. Even though several improvements were made to these guidelines, these still do not properly handle SLR appraisal. The most widely used DARE criteria assess the quality of SLRs based on a checklist that relies on the same subset of questions [22, 53, 60] identified by Kitchenham back in 2004 [25]. Unfortunately, these fail to consider whether a review is synthesized and lack appropriate search strategies. A fundamental challenge identified was the low quality of data extraction forms, review protocols, and classification schemes. These could impose ambiguity and duplicate attributes. We must separate the techniques based on their coverage of design, reporting, and evaluation [8, 34]. This is essential considering the emphasis on different stages of an SLR. Our emphasis is on the reporting. Here, the updated versions of AMSTAR, PRISMA, Kitchenham Guidelines, and CASP outperform their former versions and the other techniques.

Conducting an SLR is considered to be a highly manual, error-prone, and labor-intensive process, including tasks like data collection, extraction, and synthesis. It usually involves human judgment and decision-making. However, leveraging technologies like machine learning or natural language processing can facilitate the process. Researchers have recently developed tools to automate or semi-automate the steps or phases of the SLR process using different techniques. Previous work indicates positive results and a reduction of the effort and time required to conduct an SLR [45, 46, 48]. For instance, [10] identified several studies focused on the automation and corresponding challenges and solutions available. Similarly, automation could play a vital role in assessing the quality of SLRs efficiently and promptly. Based on the suggestions and recommendations referred to in the studies we summarize in Table 1, automation seems to be a promising direction because many steps have the potential to be automated or semi-automated.

Besides improving efficiency, automation could also enhance the quality of an assessment process by considering, for example, biases, discussions, disagreement settlements, or the synthesis process. Particularly steps in the conduct phase of an SLR are promising to automate, for instance, the search string formation, search process, and data extraction. In the reporting phase, a few steps could be automated. Since we found that most assessment techniques have pros, a valuable solution would be to modularize individual assessments to achieve a "plug and play" technique. Based on the field of research and type of review, individual modules could be configured and integrated to define the reviewing process and its assessment. For research groups, one researcher should serve as the moderator who oversees the process and ensures that the individual process steps and assessments are performed. In this context, automation would facilitate the process, assessment, and trustworthiness of an SLR by ensuring that established sets and guidelines are implemented.

6 Conclusion

In this paper, we performed a comparative analysis of prominent techniques used for quality assessing SLRs. We performed an SLR to identify these techniques. This resulted in 14 publications covering a period of 15 years (2007–2021). We identified the techniques proposed and investigated, discussed, and summarized their properties. Additionally, we could identify their strengths and weaknesses and learned that most SLRs use outdated guidelines, leading to credibility issues considering their quality, a matter that needs to be critically addressed.

We conclude that the medical domain is more advanced in literature analysis mechanisms overall, and bias detection is a major quality limitation that is still not covered in most of the studied techniques. In future work, we plan to expand our analysis by conducting an extended SLR involving other data sources, allowing us to obtain more in-depth insights and improve the validity of our work. Moreover, we plan to advance the proposed techniques regarding automation and to evaluate their validity. For this purpose, we envision to collect insights and steer discussions by conducting a workshop that invites interested parties to analyze our insights and improve them.

References

1. Centre for reviews and dissemination, about dare (2015). NHS Centre for Reviews and Dissemination (2015)
2. Critical appraisal skills programme (casp) (2018), checklist (2018), www.casp-uk.net
3. Alchokr, R., Borkar, M., Thotadarya, S., Saake, G., Leich, T.: Supporting systematic literature reviews using deep-learning-based language models. In: International Workshop on Natural Language-Based Software Engineering. p. 67–74 (2023)
4. Ali, N., Usman, M.: A critical appraisal tool for systematic literature reviews in software engineering. Information and Software Technology **112** (2019)
5. Boell, S., Cecez-Kecmanovic, D.: On being 'systematic' in literature reviews in is. Journal of Information Technology **30** (2015)
6. Bowes, D., Hall, T., Beecham, S.: SLuRp - A tool to help large complex systematic literature reviews deliver calid and rigorous results. In: International workshop of evidential assessment software technology (EAST). pp. 33–36 (2012)
7. Budgen, D., Brereton, P., Drummond, S., Williams, N.: Reporting systematic reviews: Some lessons from a tertiary study. Information and Software Technology **95**, 62–74 (2018)
8. Chandler, J., Churchill, R., Higgins, J., Lasserson, T., Tovey, D., et al.: Methodological expectations of cochrane intervention reviews. Sl: Cochrane Collaboration **3**(2), 1–14 (2016)
9. van Dinter, R., Tekinerdogan, B., Catal, C.: Automation of systematic literature reviews: A systematic literature review. Information and Software Technology **136** (2021)
10. van Dinter, R., Tekinerdogan, B., Catal, C.: Automation of systematic literature reviews: A systematic literature review. Information and Software Technology **136** (2021)
11. Dixon-Woods, M., Agarwal, S., Jones, D., Young, B., Sutton, A.: Synthesising qualitative and quantitative evidence: a review of possible methods. Journal of Health Services Research & Policy **10**(1), 45–53 (2005)
12. Donnelly, C.A., Boyd, I., Campbell, P., Craig, C., Vallance, P., Walport, M., Whitty, C.J., Woods, E., Wormald, C.: Four principles to make evidence synthesis more useful for policy. Nature **558**, 361–364 (2018)
13. Durand, G.C., Janardhana, A., Pinnecke, M., Shakeel, Y., Krüger, J., Leich, T., Saake, G.: Exploring large scholarly networks with hermes. Extending Database Technology (2018)

14. Fabbri, S., Silva, C., Hernandes, E., Octaviano, F., Di Thommazo, A., Belgamo, A.: Improvements in the start tool to better support the systematic review process. In: International conference on evaluation and assessment in software engineering (EASE)). pp. 1–5 (2016)
15. Felizardo, K.R., Carver, J.C.: Automating systematic literature review. Contemporary Empirical Methods in Software Engineering pp. 327–355 (2020)
16. Fernández-Sáez, A.M., Bocco, M.G., Romero, F.P.: SLR-Tool a tool for performing systematic literature reviews. In: ICSOFT. pp. 157–166 (2010)
17. Hernandes, E., Zamboni, A., Fabbri, S., Di Thommazo, A.: Using GQM and TAM to evaluate StArt – a tool that supports systematic review. CLEI Electron. J. **15**(1) (2012)
18. Hoffmann, F., Allers, K., Rombey, T., Helbach, J., Hoffmann, A., Mathes, T., Pieper, D.: Nearly 80 systematic reviews were published each day: observational study on trends in epidemiology and reporting over the years 2000-2019. Journal of Clinical Epidemiology **138** (2021)
19. Hong, Q.N., Fàbregues, S., Bartlett, G., Boardman, F., Cargo, M., Dagenais, P., Gagnon, M.P., Griffiths, F., Nicolau, B., O'Cathain, A., Rousseau, M.C., Vedel, I., Pluye, P.: The mixed methods appraisal tool (mmat) version 2018 for information professionals and researchers. Education for Information **34**, 1–7 (2018)
20. Jonnalagadda, S.R., Goyal, P., Huffman, M.D.: Automating data extraction in systematic reviews: a systematic review. Systematic Reviews **4**(1), 1–16 (2015)
21. JPT, H., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M., Welch, V.: Cochrane handbook for systematic reviews of interventions version 6.5. John Wiley Sons (2024)
22. Khan, A.A., Keung, J., Niazi, M., Hussain, S., Zhang, H.: Systematic literature reviews of software process improvement: A tertiary study. In: Systems, software and services process improvement. pp. 177–190 (2017)
23. Kitchenham, B.: Procedures for performing systematic reviews. Keele, UK, Keele Univ. **33** (2004)
24. Kitchenham, B., Brereton, P.: A systematic review of systematic review process research in software engineering. Information and Software Technology **55**, 2049–2075 (2013)
25. Kitchenham, B., Pretorius, R., Budgen, D., Pearl Brereton, O., Turner, M., Niazi, M., Linkman, S.: Systematic literature reviews in software engineering - a tertiary study. Inf. Softw. Technol. **52**(8), 792–805 (2010)
26. Kitchenham, B.A., Charters, S.: Guidelines for performing systematic literature reviews in software engineering. Tech. rep., Keele University and University of Durham (2007)
27. Kitchenham, B.A., Dyba, T., Jorgensen, M.: Evidence-based software engineering. In: International conference on software engineering (ICSE). pp. 273–281 (2004)
28. Kitchenham, B.A., Budgen, D., Brereton, P.: Evidence-based software engineering and systematic reviews, vol. 4 (2015)
29. Krüger, J., Lausberger, C., von Nostitz-Wallwitz, I., Saake, G., Leich, T.: Search. review. repeat? an empirical study of threats to replicating SLR searches. Empirical Software Engineering **25**(1) (2020)
30. Liberati, A., Altman, D., Tetzlaff, J., Mulrow, C., Gøtzsche, P., A. Ioannidis, J., Clarke, M., P. J. Devereaux, Kleijnen, J., Moher, D.: The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. Journal of Clinical Epidemiology **62**(10), 1–34 (2009)

31. Long, H.A., French, D.P., Brooks, J.M.: Optimising the value of the critical appraisal skills programme (casp) tool for quality appraisal in qualitative evidence synthesis. Research Methods in Medicine & Health Sciences **1**(1), 31–42 (2020)
32. Malheiros, V., Höhnr, E., Pinho, R., Mendonca, M., Maldonado, J.C.: A visual text mining approach for systematic reviews. In: International symposium on empirical software engineering and measurement (ESEM). pp. 245–254 (2007)
33. Marshall, C., Brereton, P.: Tools to support systematic literature reviews in software engineering: a mapping study. In: International conference on evaluation and assessment in software engineering (EASE) (2013)
34. Moher, D., Liberati, A., Tetzlaff, J., Altman, D.: Preferred reporting items for systematic reviews and meta-analyses: the prisma statement. Plos Med **8**, 336–341 (2009)
35. Moher, D., Cook, D., Eastwood, S., Olkin, I., Rennie, D., Stroup, D.: Improving the quality of reports of meta-analyses of randomised controlled trials: The quorom statement. Lancet **354**, 1896–900 (1999)
36. Molleri, J., Silva, L., Benitti, F.: Proposal of an automated approach to support the systematic review of literature process. In: International Conference on Software Engineering & Knowledge Engineering (SEKE). pp. 488–493 (2013)
37. Norris, R., Webb, J., Nichols, S., Stewardson, M., Harrison, E.: Analyzing cause and effect in environmental assessments: using weighted evidence from the literature. Freshwater Science **31**, 5–21 (2012)
38. Okoli, C., Schabram, K.: A guide to conducting a systematic literature review of information systems research. SSRN (2010)
39. Osborne, J.: What is rotating in exploratory factor analysis? Assessment **20**, 1–8 (2015)
40. Page, M.J., McKenzie, J.E., Bossuyt, P.M., Boutron, I., Hoffmann, T.C., Mulrow, C.D., Shamseer, L., Tetzlaff, J.M., Moher, D.: Updating guidance for reporting systematic reviews: development of the PRISMA 2020 statement. Journal of Clinical Epidemiology **134**, 103–112 (2021)
41. Paré, G., Trudel, M.C., Jaana, M., Kitsiou, S.: Synthesizing information systems knowledge: a typology of literature reviews. Information & Management **52**(2), 183–199 (2015)
42. Pejić-Bach, M., Cerpa, N.: Editorial: planning, conducting and communicating systematic literature reviews. Journal of Theoretical and Applied Electronic Commerce Research **14**(3), 190–192 (2019)
43. Ponsard, A., Escalona, F., Munzner, T.: Paperquest: A visualization tool to support literature review. In: Proceedings of the conference extended abstracts on human factors in computing systems CHI. pp. 2264–2271 (2016)
44. Ralph, P., Baltes, S.: Paving the way for mature secondary research: the seven types of literature review. In: Proceedings of the joint european software engineering conference and symposium on the foundations of software engineering FSE. pp. 1632–1636 (2022)
45. Shakeel, Y., Alchokr, R., Krüger, J., Leich, T., Saake, G.: Incorporating altmetrics to support selection and assessment of publications during literature analyses. In: International conference on vvaluation and assessment in software engineering (EASE)). p. 180–189 (2022)
46. Shakeel, Y., Krüger, J., Nostitz-Wallwitz, I.V., Saake, G., Leich, T.: Automated selection and quality assessment of primary studies: A systematic literature review. J. Data and Information Quality **12**(1) (2019)

47. Shakeel, Y., Krüger, J., Saake, G., Leich, T.: Indicating studies' quality based on open data in digital libraries. In: Business Information Systems Workshops. pp. 579–590 (2019)
48. Shakeel, Y., Krüger, J., Nostitz-Wallwitz, I.v., Lausberger, C., Durand, G.C., Saake, G., Leich, T.: Automated literature analysis - threats and experiences. In: International Workshop on Software Engineering for Science (SE4Science). pp. 20–27 (2018)
49. Shea, B., Grimshaw, J., Wells, G., Boers, M., Andersson, N., Hamel, C., Porter, A., Tugwell, P., Moher, D., Bouter, L.: Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews. BMC Medical Research Methodology **7**(10), 1471–2288 (2007)
50. Shea, B.J., Reeves, B.C., Wells, G., Thuku, M., Hamel, C., Moran, J., Moher, D., Tugwell, P., Welch, V., Kristjansson, E., Henry, D.A.: AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. BMJ **358** (2017)
51. Sherif, R., Pluye, P., Gore, G., Granikov, V., Hong, Q.N.: Performance of a mixed filter to identify relevant studies for mixed studies reviews. Journal of the Medical Library Association **104**, 47–51 (2016)
52. Shreffler J, H.M.: Common pitfalls in the research process. Treasure Island (FL): StatPearls (1) (2022)
53. Silva, F., Santos, A., Soares, S., França, C., Monteiro, C., Maciel, F.: Six years of systematic literature reviews in software engineering: an updated tertiary study. Information and Software Technology **53**, 899–913 (2011)
54. Thomas, J., Harden, A.: Methods for the thematic synthesis of qualitative research in systematic reviews. BMC Medical Research Methodology **8**, 45 (2008)
55. Tong, A., Flemming, K., McInnes, E., Oliver, S., Craig, J.: Enhancing transparency in reporting the synthesis of qualitative research: Entreq. BMC Medical Research Methodology **12**(1), 1–8 (2012)
56. Tsafnat, G., Glasziou, P., Choong, M.K., Dunn, A., Galgani, F., Coiera, E.: Systematic review automation technologies. Systematic reviews **3**, 1–15 (2014)
57. Webb, J., Miller, K., Stewardson, M., de Little, S., Nichols, S., Wealands, S.: An online database and desktop assessment software to simplify systematic reviews in environmental science. Environmental Modelling and Software **64**, 72–79 (2014)
58. Webster, J., Watson, R.T.: Analyzing the past to prepare for the future: writing a literature review. MIS Quarterly **26**(2) (2002)
59. Wohlin, C.: Guidelines for snowballing in systematic literature studies and a replication in software engineering. In: International Conference on Evaluation and Assessment in Software Engineering EASE. pp. 1–10 (2014)
60. Wohlin, C., Runeson, P., Hst, M., Ohlsson, M.C., Regnell, B., Wessln, A.: Experimentation in Software Engineering. Springer (2012)
61. Zhang, H., Babar, M.A.: Systematic reviews in software engineering: an empirical investigation **55**(7), 1341–1354 (2013)