2011 International Symposium on Empirical Software Engineering and Measurement

Identifying Strategies for Study Selection in Systematic Reviews and Maps

Kai Petersen[+,*], Nauman Bin Ali [+]
Blekinge Institute of Technology [+], Karlskrona, Sweden
Ericsson AB [*], Box 518, Karlskrona, Sweden
kai.petersen@bth.se/nauman.ali@bth.se

***Abstract*—Study selection in systematic reviews is prone to bias and there exist no commonly defined strategies of how to reduce the bias and resolve disagreement between researchers. This study aims at identifying strategies for bias reduction and disagreement resolution. A review of existing systematic reviews is conducted for study selection strategy identification. In total 13 different strategies have been identified.**

***Keywords*-Evidence based software engineering, systematic review, inclusion and exclusion, paper selection**

I. Introduction

The goal of evidence-based software engineering is the integration of evidence based on a research question in order to provide the best recommendation to practitioners to solve a problem [1]. Systematic reviews and mapping studies are a commonly used technique in evidence based medicine in order to aggregate evidence. Currently, systematic review/mapping as a method is widely used in software engineering research based on guides developed for the software engineering research field [2], [3], [4]. Systematic reviews aim at reducing bias through well defined steps included in the systematic review, also enabling replication. The steps include the search, selection, quality assessment, extraction, and aggregation of evidence.

It is important to capture relevant studies in the search without bias, and not to exclude relevant evidence during the selection of evidence. The selection of evidence is often done in several steps, starting with reading titles and abstracts, followed by quality assessment [3]. However, given that the inclusion and exclusion is done by humans there is a risk of bias, which might lead to exclusion of relevant studies, or inclusion of irrelevant ones. Bias in study selection is commonly reported in systematic reviews (see e.g. [5], [6]), and therefore disagreements occur between researchers conducting reviews together. In case the titles or abstracts are not clear, the bias is further fortified.

In order to understand the strategies used by researchers to reduce bias and resolve disagreements in systematic reviews and maps, a review of existing systematic reviews and maps in software engineering and computer science is conducted. The contribution is a set of strategies that can be followed by researchers. The benefit for research is a common understanding of existing strategies supporting informed decision making of what strategy to follow. Furthermore, it aids in reporting the strategies by making them explicit.

II. Related Work

In the first guide to conduct systematic reviews [2] different strategies have been proposed. The first strategy is to assess the goodness of the objectivity of inclusion/exclusion criteria by calculating inter-rater agreement using Cohen Kappa statistic. In addition, additional persons should be involved to discuss inclusion and exclusion, especially when the step is done by a single researcher. In case of uncertainty sensitivity analysis is proposed as a solution, but no detailed guide is given of how to conduct the sensitivity analysis.

In the updated guidelines [3] it is added that every agreement/disagreement needs to be resolved through discussion. Another recommendation for single researchers was proposed, namely test-retest. In test-retest a random sample of studies is re-evaluated by a single researcher to determine intra-rater reliability.

Overall, inclusion and exclusion is still a challenge and a common and well defined process for inclusion and exclusion has not been proposed. In fact, an interview study with researchers [7] has shown that study selection and getting agreement are two of the main challenges in systematic reviews. Hence, one of the success factors mentioned is the need of clear criteria.

To the best of our knowledge this is the first study focusing on the investigation of inclusion and exclusion strategies for systematic reviews in software engineering.

III. Methods

A. Research Questions

The aim of this study is to identify strategies for reducing bias and resolving disagreement between researchers conducting a systematic review. For this purpose following research questions was asked:

- *RQ1: What strategies are reported within systematic literature reviews in the area of software engineering and computer science?* The first question is answered through a literature review.

B. Review

This section describes the review procedure. We would like to point out that this is not a systematic review aggregating evidence. The goal is to arrive at an overview of what alternative strategies are available, but we are not able

978-0-7695-4604-9/11 $26.00 © 2011 IEEE
DOI 10.1109/ESEM.2011.46

to provide information about the accuracy of the strategies. *Study Identification:* The study identification was based on three commonly referred to guidelines for systematic reviews (cf. [2], [3]) and systematic maps [4]. Papers based on the guidelines have been selected as the guidelines have become a standard reference for systematic reviews in software engineering. Papers that are systematic, but were published before the guidelines have been released, will not be included. The studies citing the guidelines were identified through Google Scholar.

Study Selection: For the selection of literature the following inclusion criteria were defined:

- The abstract or title has to explicitly state that the article is a literature review or systematic literature review.
- The article is in the area of software engineering or computer science.
- The article is a journal paper, conference paper, thesis, or technical report. As Google Scholar is able to capture gray literature theses and technical reports are considered as well.

Articles are excluded from this study based on the following exclusion criteria:

- Article is not in English.
- The retrieved document is an editorial or an introduction to proceedings.
- The articles is not within the area of software engineering/computer science.
- The article is not accessible in full-text.
- The article is a duplicate of an article already in the set.

The inclusion and exclusion criteria are objective, and are easy to check without requiring interpretation (e.g. looking for the word literature review/systematic literature review in the title and abstract). Furthermore, there is little risk of bias with regard to the identification of the research area given that the guidelines [2], [3], [4] target software engineering and computer science researchers. As a consequence the choice was made that the first author conducts the inclusion/exclusion process individually.

Data Extraction and Analysis: A paper was only considered for data extraction when the review protocol/method section within the paper provides information of strategies for reducing bias/resolving disagreements. This information is usually found under the heading inclusion/exclusion and paper selection, or in the section "Conducting the review". In the data extraction the author names, title, and strategies for each paper were extracted. The following process was followed to identify strategies:

- *S1:* Identify the reported strategy and create a code for the strategy. Log the code in the data record for the paper currently under review.
- *S2:* Identify the next strategy and determine whether there already exist a code for that strategy. If a code exist, log the code for the paper currently being under review, otherwise create a new code and log the code.
- *S3:* Repeat step *S2* until the last paper/last strategy in the set has been recorded.

The coding was also done individually by the first author. In the case of strategy identification and documentation of strategies a threat of subjective interpretation is present (see Section III-C).

The guidelines delivered 300 hits for the 2004 version and 122 hits for the 2007 version. The mapping guidelines delivered 19 hits. After applying the inclusion/exclusion criteria 139 systematic reviews in software engineering/computer science were left. These were used as a basis for strategy identification. Papers that did not report any strategies for bias reduction and disagreement resolution were discarded. In the end of the process 40 articles containing strategies remained.

C. Validity of the Study

The main validity threats in this study are that strategies are missed, bias in the interpretation of strategies, and the use of a single search engine.

Missing Strategies: One threat to validity is that strategies for inclusion and exclusion are missed due to bias of the researcher. This threat is considered relatively low as after reviewing the first 7 of 40 papers 9 out of 13 codes/strategies have been identified and almost all strategies in the remaining papers fit well into these categories. There are two exceptions. Strategy 13 was found in the 35th [8] paper reviewed, and strategies 14, 15, and 16 were found in the 41st paper reviewed [9]. With each additional review considered the number of newly identified strategies reduced and stabilized after paper 7.

Interpretation of Strategies: The strategies were interpreted by a single researcher, which makes this step prone to bias.

Single Search Engine: The use of a single search engine leads to a risk of missing systematic reviews. We also only focused on all articles mentioning the guidelines. However, even though we might miss papers due to the limitation in search engines, we saw that most of the strategies were already identified after reading 7 of 40 papers. Hence, this threat is considered as being under control.

IV. Identified Strategies (RQ1)

In total we identified a total of 13 codes representing different strategies. The strategies are grouped according to their goals. Three goals have been identified:

- *Objectivity of Criteria:* Strategies verify the objectivity of the selection criteria.
- *Resolve Uncertainty and Disagreement:* Strategies aid researchers in resolving uncertainties and disagreements.

- *Clear Decision Rules:* Strategies based on decision rules determine whether an article is included or excluded.

In the following an overview of the goals and their related strategies is presented. Each table also contains the references to the studies applying the strategy. It is important to point out that the researchers might have used more of the presented strategies in their studies, but did not report them. However, it is still interesting to observe the number of reported articles as they represent what the researchers think are important strategies when conducting the paper selection. At the same time reporting a strategy means that an informed decision has been taken about that strategy.

Table I presents strategies related to the goal "Objectivity of Criteria". It can be seen that five studies followed strategy O1 and ten studies strategy O2. One possible reason for the frequent usage is that these strategies were recommended in the systematic review guidelines [2], [3]. Objective O1 is related to piloting the inclusion and exclusion criteria. Furthermore, O2 tests the objectivity considering the level of agreement. Objective criteria formulation was explicitly stated as a strategy in [5], the reason being that the review was conducted by a single author.

Table I
STRATEGIES OBJECTIVE CRITERIA

ID	Code	Descr.	No. of citations
O1	Objective Criteria Assessment (pre-inclusion/exclusion)	Test sub-sets of articles with two or more persons before starting the actual inclusion/exclusion process, high agreement indicate objective criteria	5
O2	Objective Criteria Assessment (post-inclusion/exclusion)	Reviewers measure their agreement after completing inclusion/exclusion to determine level of objectivity on all or a sample set of studies (can also be done on a sub-set)	10
O3	Objective Formulation of Criteria	Require objective statements (e.g. is X stated, Yes/No)	1

Table II presents strategies related to the goal "Resolve Uncertainty and Disagreement". A similar observation as for the previous goal can be made. Both strategies that are frequently reported have been proposed in [2], [3], i.e. to consult additional researchers, and to discuss and resolve uncertainties.

Table II
STRATEGIES RESOLVE DISAGREEMENTS/UNCERTAINTIES

ID	Code	Descr.	No. of citations
R1	Another person check	Additional reviewer(s) is/are consulted to support in the decision of inclusion or exclusion by reviewing/assessing the result	14
R2	Second vote on uncertain and exclude	Only for papers rated as either uncertain or exclude a second vote is obtained.	1
R3	If disagreement or uncertainty in decision then discuss	If a set of researchers is in disagreement, then a decision for the next step is taken after discussion to resolve the disagreement.	18

Table III presents strategies related to the goal "Clear Decision Rules". The decision rules are new and have not been reported in the guidelines [2], [3]. Strategy D5 is of interest, as this strategy is inclusive leading to more papers in the following steps. At the same time strategy D5 is in line with the recommendation given by Kitchenham [2], [3] advising to be inclusive in study selection. The remaining strategies were reported in individual studies. It should also be observed that reporting a single strategy does not mean that this is the only strategy applied. For example, regarding D4 we cannot be sure if the paper is also included if one reviewer says "exclude", and the other "uncertain". This further supports the claim that making strategies explicit is a prerequisite for complete reporting.

Table III
DECISION RULES TO DIRECTLY ARRIVE AT A DECISION

ID	Code	Descr.	No. of citations
D1	Majority Vote	The group of researchers take a vote on the article and the decision of the majority is followed	1
D2	At least one "include" then include	If one of the reviewers includes the paper then it is considered for being an included primary study	1
D3	All "include" then include	Only if all reviewers include the article then it is included, otherwise it is excluded	1
D4	At least one "uncertain" then include	If one of the reviewers is uncertain regarding the paper, it is considered in the next step of the systematic review	11
D5	One "exclude" and one "uncertain" then exclude	If one of the reviewers says exclude the other uncertain then the paper is excluded	1
D6	All researchers vote "uncertain" then include	If all researchers vote uncertain, then the paper is included	1
D7	All "exclude" then exclude	If all reviewers agree that the paper should be excluded then it is excluded, otherwise it is included	1

Overall, the analysis shows that the strategies that have been mentioned in the guidelines are most frequently reported. However, researchers apply strategies beyond that. In total nine strategies have been identified that are not part of the guidelines.

Combinations of strategies have been used as well, which is not visible from the tables right away. Most commonly two strategies have been reported. TOnly four studies reported the usage of more than two strategies (cf. [10], [8], [11], [9]). Our investigation revealed that the strategies proposed in the guidelines are often followed together (e.g. R3 with O2, and R1 with R3) [2], [3].

It is also visible that many different strategies have not been reported together at all. This raises two open questions, namely:

- *Question 1:* Which strategies to combine to get high effectiveness in selection (selecting papers relevant for the population and not excluding relevant papers) and efficiency (reduce the effort in paper selection and subsequent review steps)?
- *Question 2:* Which process should be followed (order

in which strategies are executed), e.g. at which stage of the process should we follow decision rules, discuss, or calculate inter-rater agreement to achieve effectiveness and efficiency?

V. Discussion

One aim of the systematic reviews is to follow a repeatable process [3]. Therefore it is imperative that the selection criteria and steps to resolve disagreements are documented and reported in systematic reviews. Doing this will also show that a conscious decision was made from the different available choices and bring more transparency in the review process. Furthermore, the availability of guidelines leads to adoption and reporting, as in the strategy identification review it was clearly visible that the guidelines reported in [2], [3] were the ones most frequently adopted. It was also noticed that the decisions rules in the reviews did not clearly explain how the various possibilities were handled, e.g. 8 studies use *D4* which states to include an article if there is at least one "uncertain" (as shown in Table III). It is not clear whether it means we include the studies even if two of the three reviewers classified it as irrelevant. If it does, is it a justified choice considering the effort that will be put again to review something that might very well be irrelevant.

As mentioned earlier, the reason for disagreements is that abstracts are often not clear. In general, given the lack of clarity one has to look at additional information and the article should not be excluded too early. For example, an exclusive strategy would be D5 while a more inclusive strategy would be D4. From a validity point of view D4 would strengthen the study, but would require more effort reading additional parts of the article. With a very high number of articles this effort could mean that the review is not manageable in reasonable time and becomes outdated during writing. With a lower number of articles the more inclusive strategy would be preferable. In short, knowing about the strategies helps in making informed research design decisions with respect to inclusion and exclusion.

VI. Conclusion

This study targets the problem of bias in selection studies for systematic reviews and maps (inclusion and exclusion). From existing systematic reviews strategies are identified aiming at answering the following research question: *RQ1: What strategies are reported within systematic literature reviews in the area of software engineering and computer science?* Thirteen different strategies for inclusion and exclusion have been identified. Three are used to assure objective inclusion/exclusion criteria to reduce bias, three to resolve disagreements and uncertainties due to bias, and seven defined decision rules on how to act based on disagreements/agreements.

VII. Acknowledgement

This work was funded by ELLIT. Thanks also go to Claes Wohlin for proof reading.

References

[1] B. A. Kitchenham, T. Dybå, and M. Jørgensen, "Evidence-based software engineering," in *Proceedings of the 26th International Conference on Software Engineering (ICSE 2004)*, 2004, pp. 273–281.

[2] B. Kitchenham, "Procedures for performing systematic reviews," Department of Computer Science, Keele University, ST5 5BG, UK, Tech. Rep. TR/SE-0401, 2004.

[3] ——, "Guidelines for performing systematic literature reviews in software engineering," Department of Computer Science, Keele University, ST5 5BG, UK, Tech. Rep. EBSE-2007-01, 2007.

[4] K. Petersen, R. Feldt, S. Mujtaba, and M. Mattsson, "Systematic mapping studies in software engineering," in *Proceedings of the 12th International Conference on Evaluation and Assessment in Software Engineering (EASE 2008)*, 2008.

[5] K. Petersen, "Measuring and predicting software productivity: A systematic map and review," *Information & Software Technology*, vol. 53, no. 4, pp. 317–343, 2011.

[6] R. Prikladnicki and J. L. N. Audy, "Process models in the practice of distributed software development: A systematic review of the literature," *Information & Software Technology*, vol. 52, no. 8, pp. 779–791, 2010.

[7] M. A. Babar and H. Zhang, "Systematic literature reviews in software engineering: Preliminary results from interviews with researchers," in *Proceedings of the Third International Symposium on Empirical Software Engineering and Measurement (ESEM 2009)*, 2009, pp. 346–355.

[8] W. Afzal, R. Torkar, and R. Feldt, "A systematic review of search-based testing for non-functional system properties," *Information & Software Technology*, vol. 51, no. 6, pp. 957–976, 2009.

[9] S. Jalali and C. Wohlin, "Agile practices in global software engineering - a systematic map," in *Proceedings of the International Conference on Global Software Engineering (ICGSE 2010)*, 2010, pp. 45–54.

[10] A. Meneely, B. Smith, and L. Williams, "Software metrics validation criteria: A systematic literature review," North Carolina State University Department of Computer Science, Raleigh, NC 27695-8206 USA, Tech. Rep. TR-2010-2, 2010.

[11] N. Condori-Fernández, M. Daneva, K. Sikkel, R. Wieringa, Ó. D. Tubío, and O. Pastor, "A systematic mapping study on empirical evaluation of software requirements specifications techniques," in *Proceedings of the Third International Symposium on Empirical Software Engineering and Measurement (ESEM 2009)*, 2009, pp. 502–505.