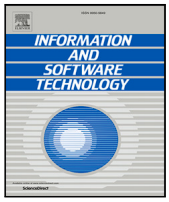
Information and Software Technology 158 (2023) 107174

Contents lists available at ScienceDirect

Information and Software Technology

journal homepage: www.elsevier.com/locate/infsof



Double-counting in software engineering tertiary studies — An overlooked threat to validity

Jürgen Börstler [a,*], Nauman bin Ali [a], Kai Petersen [a,b]

[a] Department of Software Engineering, Blekinge Institute of Technology, Sweden
[b] University of Applied Sciences Flensburg, Germany

ARTICLE INFO

Keywords:
Bias
Double-counting
Empirical
Guidelines
Meta-review
Overview of reviews
Recommendations
Research methods
Review of reviews
Tertiary review
Tertiary study
Umbrella review

ABSTRACT

Context: Double-counting in a literature review occurs when the same data, population, or evidence is erroneously counted multiple times during synthesis. Detecting and mitigating the threat of double-counting is particularly challenging in tertiary studies. Although this topic has received much attention in the health sciences, it seems to have been overlooked in software engineering.
Objective: We describe issues with double-counting in tertiary studies, investigate the prevalence of the issue in software engineering, and propose ways to identify and address the issue.
Method: We analyze 47 tertiary studies in software engineering to investigate in which ways they address double-counting and whether double-counting might be a threat to validity in them.
Results: In 19 of the 47 tertiary studies, double-counting might bias their results. Of those 19 tertiary studies, only 5 consider double-counting a threat to their validity, and 7 suggest strategies to address the issue. Overall, only 9 of the 47 tertiary studies, acknowledge double-counting as a potential general threat to validity for tertiary studies.
Conclusions: Double-counting is an overlooked issue in tertiary studies in software engineering, and existing design and evaluation guidelines do not address it sufficiently. Therefore, we propose recommendations that may help to identify and mitigate double-counting in tertiary studies.

1. Introduction

With an increasing number of systematic reviews in software engineering, tertiary studies have been published to organize or synthesize their results [1,2]. Tertiary studies represent a high level of aggregation of evidence and are, therefore, a good starting point for information about a field.[1] They can potentially reveal conflicting or confirming evidence and provide a comprehensive overview of a research topic. When the same evidence is directly or indirectly included multiple times in a tertiary study, it might be double-counted and overemphasized in the results of the tertiary study. Double-counting might therefore affect the validity and trustworthiness of the results presented in tertiary studies.

To the best of our knowledge, there are no specific guidelines for tertiary studies in software engineering. In their seminal guidelines for systematic literature reviews in software engineering, Kitchenham and Charters [3] define a tertiary study as "[a] systematic review of systematic reviews, in order to answer wider research questions" that "uses exactly the same methodology as a standard systematic literature review". However, in our experience of conducting tertiary studies [4,5], we found that several decisions and concerns differ slightly when reviewing secondary studies instead of primary studies. Specifically, issues with double-counting the evidence in primary studies when synthesizing the results from secondary studies may be easily overlooked.

Regarding double-counting in secondary studies, Kitchenham and Charters [3] note that "[i]t is important not to include multiple publications of the same data in a systematic review synthesis because duplicate reports would seriously bias any results. . . . When there are duplicate publications, the most complete should be used".

* Corresponding author.
E-mail addresses: jurgen.borstler@bth.se (J. Börstler), nauman.ali@bth.se (N. bin Ali), kai.petersen@bth.se, kai.petersen@hs-flensburg.de (K. Petersen).

[1] In 2021 alone, 16 tertiary studies in software engineering have been published according to a SCOPUS search on Nov 8, 2022, using search string *TITLE-ABS-KEY ("tertiary study" OR "tertiary review" OR "review of reviews") AND (LIMIT-TO (PUBYEAR, 2021)) AND (LIMIT-TO (SUBJAREA, "COMP")) AND (LIMIT-TO (LANGUAGE, "English"))* followed by a screening of titles and abstracts to identify tertiary studies within software engineering.

https://doi.org/10.1016/j.infsof.2023.107174
Received 31 August 2022; Received in revised form 9 January 2023; Accepted 9 February 2023
Available online 23 February 2023

Table 1
Types of double-counting in tertiary studies.

		Causes of double counting		
		Duplication	Redundancy	Overlap
Sources for double-counting	Secondary study	*Duplicate secondary studies*	*Redundant secondary studies*	*Overlap of primary studies*
	Primary study	*Duplicate primary studies*	*Redundant primary studies*	*Overlap of primary data*

Double-counting and the consequence of overstating evidence in tertiary studies have received a lot of attention in the health sciences [6–8].[2] However, it seems that authors of tertiary studies in software engineering, including ourselves, have not extended this advice sufficiently to the analysis and synthesis of secondary studies and are mostly content with looking at duplicate publications. Mitigating double-counting in a tertiary study can be more complex than identifying the most complete version of a secondary study.

In Table 1, we summarize the main types of double-counting relevant for tertiary studies and discuss them in more detail in Sections 2 and 9.

In this paper, we bring attention to the currently overlooked threat of double-counting primary studies (Overlap of primary studies in Table 1). Another potential threat is the double-counting of data (Overlap of primary data in Table 1). However, we have not analyzed the tertiary studies for overlaps in primary data in detail.

The main contributions of this paper are as follows:

- A discussion and exemplification of double-counting issues in tertiary studies in software engineering.
- An analysis of the prevalence of double-counting issues in tertiary studies in software engineering and how they have been addressed.
- A list of recommendations for tertiary studies in software engineering.

In this paper, we first discuss potential causes for double-counting in tertiary studies according to Table 1 in Section 2 and then discuss the overlap of primary studies in more detail in Section 3. The related work and the method used in our study are described in Section 4 and Section 5, respectively. Thereafter, we analyze 47 tertiary studies in software engineering to assess if double-counting is recognized as an issue and which mitigation strategies are used to address it (Section 6). Based on the results (Section 6) and their analysis (Section 7), we propose recommendations for future tertiary studies (Section 9). In Section 10, we demonstrate how our proposed recommendations would have helped to identify and mitigate double-counting threats in our sample tertiary studies. Section 11 concludes the paper.

2. Potential causes for double-counting in tertiary studies

Duplicate primary/secondary studies. A duplicate study is a "literal" duplicate of another study. Such duplicates may be the result of multiple occurrences of the same study found using different searches, e.g., due to finding the same study using different search engines or different search strategies. A duplicate may also result from an indexing error in a literature database or slight differences in the metadata in the same or different databases. Another source of duplication is republication.

It should be noted that duplicates may have different DOIs (e.g., in the case of republication). Comparing DOIs is, therefore, not a fully reliable approach to identifying duplicates.

Redundant primary/secondary studies. A redundant study is a study that has been replaced or superseded by another study that is not a duplicate. Redundant studies may result from extending, updating, or replacing a study with a (typically) newer and/or more comprehensive version. A redundant study can, for example, be a conference publication extended to a journal publication, an update or extension of an existing study (e.g., by changing its coverage or time-frame), or a technical or self-archived report that has been published formally in a peer-reviewed venue.

It should be noted that identifying redundant studies may be difficult since studies might not discuss relationships with other studies thoroughly. In a tertiary study, Verner et al. [9], for example, point out that "SLRs are supposed to comment on other SLRs covering the same or related material. However, most of the SLRs we reviewed do not reference related SLRs and so do not define their overlap with other SLRs". Similar observations have been made for primary studies [10].

Overlap of primary studies. A root cause for double-counting in tertiary studies is an overlap of the primary studies in the included secondary studies. An included secondary study may include a primary study (duplicate or redundant) that is also included in one or more other secondary studies. If this overlap is not considered, the evidence presented in the duplicate and redundant primary studies might be overemphasized in the tertiary study. This issue is discussed and exemplified in more detail in Section 3.

Overlap of primary data. An overlap of primary data exists when multiple primary studies use the same primary data, such as public datasets, systems, cases, or populations. This may lead to an over-representation of those data in secondary studies. If this overlap is not considered when conducting a secondary study, the evidence related to the overlapping primary data might be overemphasized in the secondary study (and propagate to tertiary studies, including the secondary study). It should be noted, though, that even if this overlap is considered in the secondary study, it needs to be reconsidered in a tertiary study since the primary data may originate from primary studies included in different secondary studies.

An overlap of primary data may occur, for example, when primary studies use the same benchmark data (e.g., the PROMISE dataset [11] or the Software-artifact Infrastructure Repository (SIR) [12,13]), the same frequently used open-source systems, or other open sources (like GitHub and Stackoverflow). Other sources for overlaps may be the reuse of case contexts or survey/ experiment participants.

It should be noted that replications are also a potential source for the overlap of primary data. When there are only a few key primary studies that have been replicated many times, their context information may bias analyses and syntheses that are based on this information. Cruz et al.'s [14] mapping of replications in empirical software engineering further indicates that few author networks dominate the area, which might lead to bias in secondary studies that are not aware of double-counting.

[2] To the best of our knowledge, double-counting has not been dealt with outside the medical/health sciences. A SCOPUS search on Nov 3, 2022, using search string *(TITLE-ABS-KEY ((study W/1 overlap) OR (double W/1 counting)) AND (TITLE-ABS-KEY (tertiary OR mapping OR (systematic W/1 review) OR (meta W/1 analysis)))) AND (EXCLUDE (SUBJAREA, "MEDI") OR EXCLUDE (SUBJAREA, "BIOC") EXCLUDE (SUBJAREA, "PSYC") EXCLUDE (SUBJAREA, "NEUR"))* returned 44 documents. Of those 44, only one was relevant and covered a topic that is discussed in our paper.

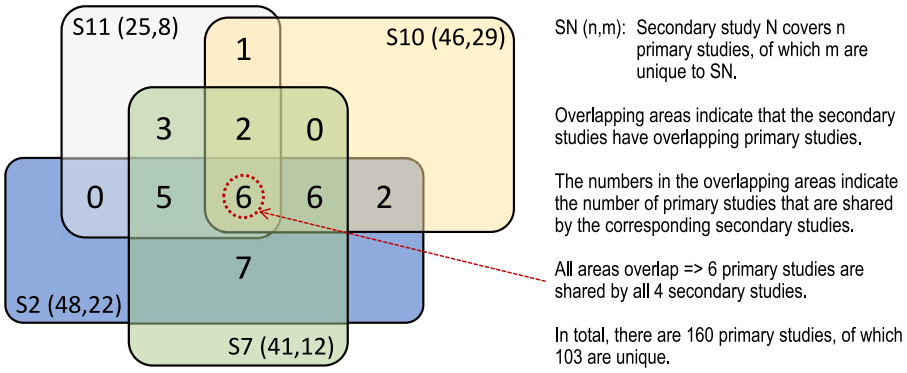


Fig. 1. Overlap of primary studies between four secondary studies in TDD, according to Nurdiani et al. [4]. For a better overview, we have complemented the presentation with the number of unique primary studies for each secondary study.

Table 2
Quality characteristics and key findings of the shared primary studies (P1–P6) as described in S7 [15].

ID	Relevance	Rigor	TDD positive	TDD no difference	TDD negative
P1	high	low	external quality		time/effort
P2	high	low			performance/productivity
P3	low	high		productivity, external quality	
P4	low	high	effort/time	productivity, internal code quality	
P5	low	high	effort/time, conformance	external quality, internal code quality	
P6	low	high	external quality	productivity, effort/time	

3. Overlap of primary studies: An example

When aggregating the results of secondary studies, one can usually not assume that the secondary studies have disjoint sets of primary studies. A tertiary study on agile practices [4], for example, found a substantial overlap of primary studies between the secondary studies dealing with Test Driven Development (TDD). As shown in Fig. 1, six primary studies are shared by the four secondary studies S2, S7, S10, and S11. These six shared primary studies might shape the synthesis of all four secondary studies and therefore affect the validity of a tertiary study that includes those secondary studies. Authors of tertiary studies need to take such overlaps of primary studies into account to avoid potential bias when synthesizing the results of secondary studies.

Nurdiani et al.'s [4] tertiary study on agile practices identified[3] six primary studies that were shared by all four of the included secondary studies about Test Driven Development (TDD). Fig. 1 depicts the overlap[4] and Table 2 summarizes the quality characteristics and main findings of the six shared primary studies. Looking at the shared primary studies from Fig. 1 in more detail reveals the following potential biases when synthesizing the results using vote counting.

- *Biases concerning research quality:* Of the six shared primary studies in [4], Munir et al. (S7 [15]) assessed them as either low rigor or low relevance. At the same time, Munir et al. assessed nine of their remaining primary studies as having high rigor and high relevance. By not taking the overlap of primary studies into consideration, studies with low rigor or low relevance might be overemphasized in Nurdiana et al.'s synthesis of the four secondary studies.
- *Biases concerning research results:* The six shared primary studies were mostly inconclusive concerning the observed variables, e.g., external quality. In the nine primary studies of high rigor and high relevance, only positive results concerning external quality were reported. This may lead to undesired biases when simply aggregating results without normalizing concerning overlapping primary studies.

Biases concerning quality and results may have an undesired interaction effect. That is, if high-quality and low-quality studies had the same distribution of vote counts, it would not be too problematic as only the effect will be overemphasized. However, in most cases, it is essential to emphasize high-quality studies over low-quality studies. Details concerning the example are available in the supplementary material (https://tinyurl.com/double-counting-in-TS).

4. Related work

As indicated in the introduction, to the best of our knowledge, there are no specific guidelines or recommendations for tertiary studies in software engineering. In secondary studies, it is common practice to delete duplicate publications, and most guidelines and recommendations extend this practice to mean "publications of the same data". In Ampatzoglou et al.'s [16] comprehensive review on validity threats in secondary studies in software engineering, the authors recommend a "consistent strategy (e.g., keep the newer one or keep the journal version) for selecting which study should be retained". Furthermore, they recommend "summaries of candidate primary studies to guarantee the correct identification of all duplicate articles". Whether or not updated or extended studies should be considered duplicates is unclear, though.

In a recent systematic mapping on tertiary studies to analyze how tertiary studies define and apply inclusion and exclusion criteria of secondary studies, Costal et al. [17] noted that 19 of 50 tertiary studies used duplication in terms of "reported in different documents" as a selection criterion. They also pointed out that the concept of duplicates is used ambiguously and could refer to what we define as duplicate

[3] The full list of the six primary and four secondary studies (S2, S7, S10, S11) can be found in Appendix A.

[4] Fig. 1 depicts overlaps according to Figure 3 in [4]. In Section 7 (Fig. 3), we present a more general and compact notation for depicting overlaps of primary studies.

or redundant, respectively, in Table 1. Other forms of duplication or double-counting were not mentioned in Costal et al.'s study, though.

This does not mean that authors of secondary or tertiary studies are not aware of the potential threats to validity that double-counting might cause. In a discussion about a secondary study on perspective-based reading, Kitchenham et al. [18, p 22], for example, noted that the review included many replications. However, they also noted that similar results were found in an included independent study. Although not explicitly mentioning double-counting, Rios et al.'s tertiary study on technical debt [19] avoided double-counting of overlapping primary studies by mapping data directly to the corresponding primary study to avoid counting a primary study multiple times in case several secondary studies share it.

In the health sciences, double-counting is discussed more explicitly and more thoroughly. A systematic review on tertiary studies[5] published 2009–2011 in the health sciences, found that "[o]nly 32 of 60 overviews mentioned overlaps" [6]. In a recent scoping review, Gates et al. [21] found 77 guidance documents for conducting overviews of reviews in the health sciences. Six of those provide "diverse guidance about how best to manage overlapping and/or discordant systematic reviews". Five of those six recommend that the "[a]uthors may decide to include all systematic reviews regardless of overlap, or only include the most recent, most comprehensive, most relevant, or highest quality systematic reviews".

The Cochrane handbook [20] contains a separate subsection on managing overlapping systematic reviews in overviews of reviews. The main advice that is transferable to tertiary studies in software engineering is to assess the overlap of primary studies. Although there is a long tradition of conducting secondary and tertiary studies in the health sciences, a recent study [7] concludes "that there is currently no standard methodological approach to deal with an overlap in primary studies across reviews".

5. Research method

To investigate the potential threat of double-counting in tertiary studies in software engineering, we posed the following research questions:

RQ1: How mindful of double-counting issues are tertiary studies in software engineering?

RQ2: Which types of double-counting issues have they identified?

RQ3: Were double-counting issues mitigated sufficiently?

RQ4: Which strategies have they used to address double-counting issues?

To answer the research questions, we capitalized on Costal et al.'s recent reviews of tertiary studies [17,22]. They investigated how tertiary studies in software engineering, published in English 2004–early 2021, perform study selection and quality assessment of the included secondary studies, respectively. We leverage their search and selection results as they are aligned with our research questions[6]. The data we extracted from the 50 tertiary studies selected by Costal et al. are summarized in Table 3. Table B.6 in Appendix B also indicates the mapping of extracted information to the research questions.

The first and second authors piloted the data extraction to reach a common interpretation of the criteria. We noted that some tertiary studies address double-counting without acknowledging it as a threat to tertiary studies in general (item #7). We, therefore, agreed to change "no"s for item #7 to "yes, implicitly" when authors consider double-counting a threat for their own tertiary study (i.e., "yes" for item #8) or when authors provide a strategy for addressing double-counting (entry for item #12).

The first and second authors then extracted the data from 25 tertiary studies each for all fields, except #6, #13, and #14. For the two tertiary studies with conflicts of interest (T15, T50), the data were extracted by the author without a conflict. For seven studies, the data could not be extracted unambiguously. Both authors discussed those studies, and the questions were resolved consensually. The first author then extracted information for the remaining fields (#6, #13, #14). Finally, the third author reviewed and validated the extracted data for all 50 studies except T50 (due to a conflict of interest). The data extraction for T50 was validated by an independent researcher.

We excluded three tertiary studies from Costal et al.'s dataset [23] during the data extraction. Two tertiary studies turned out to be hybrids between secondary and tertiary studies (T08, T39), and a third (T30) turned out to be redundant to a more recent and more complete tertiary study (T11). T08, T30, and T39 were, therefore, excluded from our dataset resulting in a total of 47 tertiary studies. We have kept Costal et al.'s original study IDs for easier cross-reference.

The full list of tertiary studies can be found in Table C.7 in Appendix C.

6. Results

In Section 6, we first present some raw data and then answer our research questions in isolation. In Section 7, we then give a visual overview of the results and discuss them in more detail.

Table 4 provides an overview of the quantitative data from our data extraction. From Costal et al. [23], we already know that the scope of the tertiary studies is roughly evenly distributed between studies investigating specific software engineering topics (26 tertiary studies) and studies investigating methodological issues of secondary studies (21 tertiary studies). Of the 47 tertiary studies, 33 conducted and reported a quality assessment of the included secondary studies. For the remaining 14, 4 explicitly stated that they did not conduct a quality assessment, and for 10 it is unknown whether they conducted one.

A comprehensive overview of the data extraction for all 47 tertiary studies is available in an electronic supplement (https://tinyurl.com/double-counting-in-TS).

6.1. RQ1: How mindful of double-counting issues are tertiary studies in software engineering?

Of the 47 tertiary studies, 9 acknowledge double-counting as a threat (implicitly or explicitly) to the validity of tertiary studies in general. Of those nine, five do also consider it a threat to their own study's validity (T04, T06, T19, T42, T43). These five are also in agreement with our assessment of the threat. Of the four studies that do not consider double-counting a threat to their validity (T01, T09, T10, T15), we consider it a threat for all four.

Of the remaining 38 tertiary studies that do not acknowledge double-counting as a threat to the validity of tertiary studies in general, we consider that double-counting is a concern for 10 of them (T02, T12, T13, T14, T17, T23, T24, T25, T38, T48). None of those 10 handles the threat sufficiently.

6.2. RQ2: Which types of double-counting issues have they identified?

Duplicate secondary studies: Of the 47 tertiary studies, 34 described that they dealt with duplicate secondary studies in some form. For 14 of those 34, it was, however, not clear whether they referred to (literal) duplicates and/or redundant secondary studies.

[5] In the health sciences the following terms are used interchangeably for tertiary studies [20]: overviews of reviews (or just overviews), umbrella reviews, reviews of reviews and meta-reviews.

[6] We used Costal et al.'s replication package [23].

Table 3
Data extracted from the tertiary studies.

#	Field
1	Unique study ID[a]
2	Scope of the tertiary study (methodological or specific SE area)[a]
3	Area of the tertiary study (e.g., software reuse, testing, search)[a]
4	Number of secondary studies included in the tertiary study[a]
5	Quality assessment of the included secondary studies (reported, not done, unknown)[a]
6	Study builds on other tertiary studies (we extracted the tertiary studies a tertiary study builds on and in which way these tertiary studies were used)
7	Study acknowledges some form of double-counting as a validity threat to tertiary studies in general (yes/yes, implicitly/no)
8	Study considers double-counting beyond duplicate/redundant secondary studies a threat for itself (yes/no)
9	Double-counting beyond duplicate/redundant secondary studies is a threat to validity for the study according to our assessment (yes/no; based on the data extracted in #11)
10	Double-counting beyond duplicate/redundant secondary studies is handled sufficiently to mitigate threats to validity according to our assessment (yes/no/not applicable since it is no threat; based on the data extracted in #11)
11	Information provided in the study regarding the handling of double-counting wrt study validity (used to answer items #9 and #10; we extracted relevant references and pointers to a study's text for further analysis)
12	Study's strategy for addressing double-counting beyond duplicate/redundant secondary studies (when the study provided a strategy, we either provided a short summary of the strategy in our own words or extracted relevant references to the study's text for further analysis)
13	Study describes the handling of duplicate/redundant secondary studies (yes, mentions duplicate and redundant publications/yes, mentions only duplicates/yes, mentions only redundant/yes but unclear which/no description/relies on single existing dataset)
14	Study justifies its need in relation to existing tertiary studies (yes /no/claims there are none)
15	Further comments

[a]According to Costal et al. [23]. We have kept their study IDs for easier cross-reference.

Table 4
Quantitative results for data extraction items #6, #7, #8, #9, #10, #12, #13.

Data extraction item	Count	Tertiary studies
Study builds on other TSs[a] (#6)	8	T05, T09, T23, T37, T44, T45, T46, T48
Study acknowledges some form of DC[b] as a validity threat to TSs in general (#7)	9	T01, T04, T06, T09, T10, T15, T19, T42, T43
Study considers DC beyond duplicate/redundant SS[c] a threat for itself (#8)	5	T04, T06, T19, T42, T43
DC beyond duplicate/redundant SS is a threat to validity for the study according to our assessment (#9)	19	T01, T02, T04, T06, T09, T10, T12, T13, T14, T15, T17, T19, T23, T24, T25, T38, T42, T43, T48
DC beyond duplicate/redundant SS is handled sufficiently to mitigate threats to validity according to our assessment (#10)	3	T04, T10, T43
Study's strategy for addressing DC beyond duplicate/redundant SS (#12)	7	T04, T06, T09, T10, T15, T42, T43
Study describes the handling of duplicate and redundant SS (#13)	17	T02, T04, T06, T13, T14, T20, T25, T29, T31, T32, T34, T36, T42, T43, T44, T47, T48

[a]TS = tertiary study.
[b]DC = double-counting.
[c]SS = secondary study.

Redundant secondary studies: Of the 47 tertiary studies, 17 explicitly describe that they deleted redundant secondary studies, like a conference publication extended to a journal publication. All 17 also noted that they deleted duplicates. One study (T42) notes that they "found 24 SLR studies reported in 37 papers" and explicitly marked redundant secondary studies in their list of secondary studies. T48 notes that two SLRs using the same dataset were excluded.

Overlap of primary studies: Of the 47 tertiary studies, 9 (T01, T03, T04, T06, T09, T10, T15, T42, T43) acknowledge double-counting of primary studies in the included secondary studies included in a tertiary study as an issue, in general, and 5 of those consider it a threat for their own validity (T04, T06, T19, T42, T43).

Overlap of primary data: Of the 47 tertiary studies, 2 (T04, T42) identify some form of double-counting of data or evidence. T42 acknowledges the problem of multiple primary studies using "the same company's participants so may not be independent". In T04, the authors map certain data items directly to the primary studies included in the secondary studies to avoid counting them multiple times via the secondary studies.

As we stated in the introduction (see also Table 1), we also consider that the reuse of underlying cases, like systems, datasets, or benchmarks in multiple primary studies, might bias the results of a secondary and a tertiary study.

6.3. RQ3: Were double-counting issues mitigated sufficiently?

Duplicate and redundant secondary studies: As described in Section 6.2, 34 tertiary studies described that they deleted duplicate and/or redundant secondary studies. From the information available in the tertiary studies, it was not possible to assess whether they mitigated these threats sufficiently. Furthermore, it should be noted that this threat still might have been sufficiently mitigated even when the deletion of duplicate and redundant secondary studies is not explicitly mentioned. Regarding redundant secondary studies, only T42 explicitly

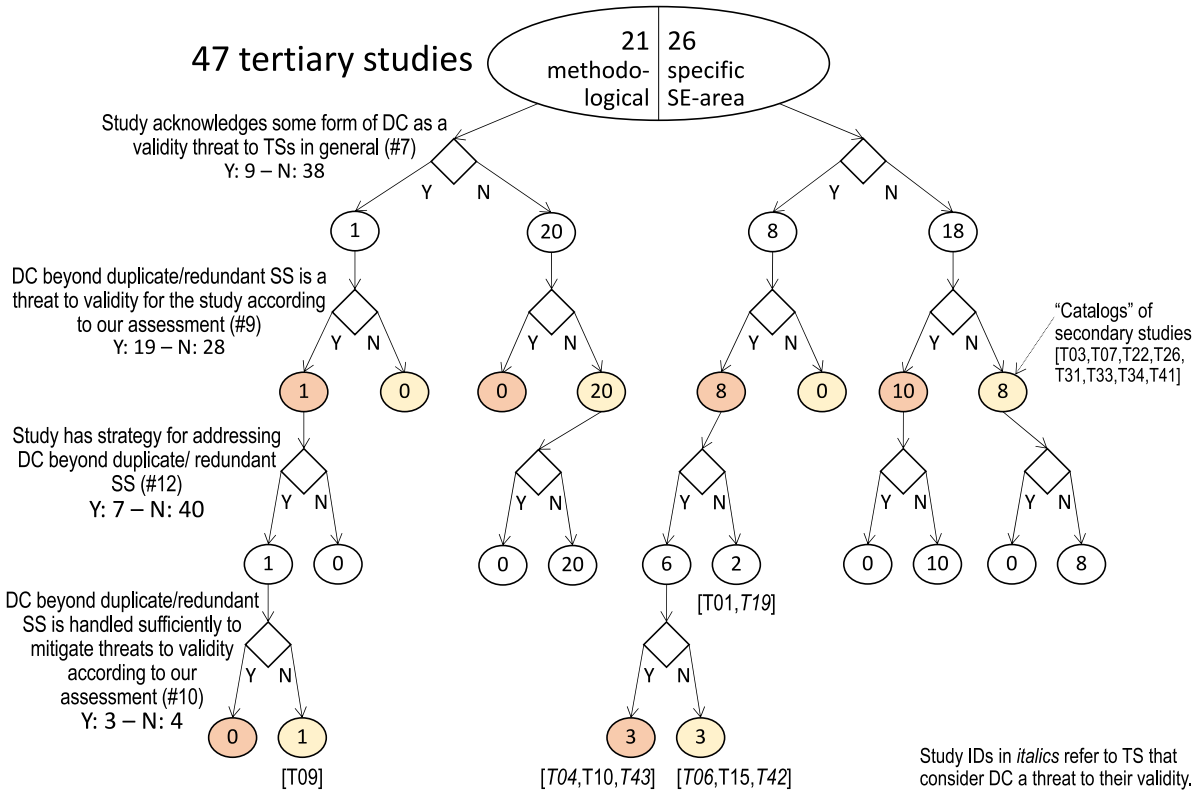


Fig. 2. Graphical representation of answers to data extraction items #7, #9, #10 and #12. T01–T50 refer to the study IDs in Appendix C.

marked redundant secondary studies and excluded them from the synthesis.

Overlap of primary studies: Of the 19 tertiary studies that we assessed to have a threat to validity beyond duplicate/redundant secondary studies, seven address the overlap of primary studies (T04, T06, T09, T10, T15, T42, T43). Of these seven, three mitigate the threat sufficiently (T04, T10, T43). It can be noted that T10 mitigates the threat, although it does not explicitly consider double-counting beyond duplicate/redundant secondary studies as a threat to its validity.

Overlap of data: Only one study in our sample (T04) mitigated an overlap of data.

6.4. RQ4: Which strategies have they used to address double-counting issues?

Seven of the 47 tertiary studies in our sample provide or suggest strategies for addressing double-counting (T04, T06, T09, T10, T15, T42, T43). Four of those seven consider double-counting a threat to their study (T04, T06, T42, T43).

Six of the seven tertiary studies explicitly discuss overlaps of primary studies. Of those six, four analyze the overlaps of primary studies (T10, T15, T42, T43), and two present graphical overviews of their analyses (T15, T43). Two of these six tertiary studies (T06 and T09) do not conduct an explicit analysis of the overlaps but state that the overlap is likely small and will not affect their studies' results.

The remaining tertiary study (T04) goes directly to the primary studies included in the secondary studies to avoid double-counting, without first analyzing overlaps. It can be noted that T04 also presents a full list of all primary studies.

6.5. Dependencies between tertiary studies

Of the 47 tertiary studies, 30 justified their need in relation to existing tertiary studies. Among those 30 were all eight tertiary studies that depend on other tertiary studies. Two of those eight (T46, T44) are extensions of T47. The remaining six tertiary studies (T05, T09, T23, T37, T45, T48) reuse (and sometimes combine) the sets of selected secondary studies from other tertiary studies. All eight dependent tertiary studies discuss their relationships to the original studies in detail.

Of the 17 tertiary studies that did not explicitly discuss their relationship to existing tertiary studies, six just state that there are no related tertiary studies (T01, T06, T10, T17, T18, T38) and one is considered the first tertiary study in software engineering (T47). The remaining ten tertiary studies do not mention related tertiary studies at all.

7. Analysis and discussion

For a large number of tertiary studies, we found that double-counting is no validity threat (28 of 47 studies). There are two main reasons for this: (a) an overwhelming number of these tertiary studies focus on methodological concerns[7] related to the conduction of secondary studies (21 of 28 studies), and (b) the remaining ones are "catalogs"[8] of secondary studies on a software engineering topic (8 of 28 studies).

In both cases, because of the aims of the studies, an analysis of evidence and research aggregated in the identified secondary studies is not of concern. Thus, the overlap of primary studies is irrelevant to these studies. These studies have only to ensure that duplicate and redundant secondary studies are deleted.

However, from Fig. 2, we can see that in our dataset overlap of primary studies is a threat for one (T09) of the 21 tertiary studies with

[7] According to Costal et al.'s categorization [22], these are tertiary studies "that focus on the methods and protocols followed by secondary studies in their development process".

[8] By "catalog", we refer to a tertiary study that collects and organizes information about secondary studies.

	S2	S7	S10	S11
S2		36.9%	17.5%	17.2%
S7			19.2%	32%
S10				14.5%
S11				

CCA = N – r / (c – 1) · r
= 160 – 103 / 3 · 103
≈ 0.184
18.4% ⇒ very high overlap

	S2	S7	S10	S11
P1	x	x	x	x
P2	x	x	x	x
P3	x	x	x	x
P4	x	x	x	x
P5	x	x	x	x
...				
P102				x
P103				x
∑	48	41	46	25

= 160

CCA (as %)	Overlap
= 0	no
0 < CCA < 6	slight
6 ≤ CCA ≤ 10	moderate
10 < CCA ≤ 15	high
> 15	very high

Fig. 3. Pairwise CCA-matrix (left) and citation matrix (middle) for our example in Section 3. The calculation for the total CCA (bottom left) is according to the numbers of primary and secondary studies, as well as overlaps, in Fig. 1. The interpretation of CCA (right) is adopted from Pieper et al. [6].

a methodological focus. This tertiary study also addresses the threat. One could argue, though, whether T09 is a methodological study. The authors of T09 write that they "are particularly interested ... in what context, and by whom, the core tasks of the primary studies were performed" (p. 236). Since the tasks refer to software engineering tasks, we argue that T09 is a combination of a methodological study and a study on software engineering topics.

Of the eight "catalogs" (T03, T07, T22, T26, T31, T33, T34, T41) double-counting beyond duplicate/redundant secondary studies is no threat to the validity for any of them.

Currently, tertiary studies are mainly used to give overviews of research areas. Overviews are necessary, but more thorough syntheses of secondary studies would be even more beneficial to advance the software engineering body of knowledge. However, basing a synthesis of a tertiary study on only the results provided in secondary studies is very difficult since these results may be aggregations/syntheses of potentially overlapping primary studies.

In our sample, four tertiary studies analyzed the overlaps of primary studies at least partially (T10, T15, T42, T43) and two presented graphical overviews of their analyses of overlaps of primary studies (T15, T43). While T43 deleted all overlaps before analyzing its research questions, T15 did not consider the overlap of primary studies a threat to its validity. T10 reported finding only four primary studies that were cited by more than one of the four included secondary studies but did not follow up on that. T42 noted that "the primary papers referred to by the SLRs in many cases overlapped" which "made it difficult to be sure about the real degree of empirical support for many items". In addition, T42 provides tables with the extracted raw data for further analysis.

In the health sciences, it is recommended to analyze the overlap of primary studies [20]. Fig. 1 in Section 3 shows an example of a comprehensive and compact overview of an overlap of primary studies. Such a presentation can become unwieldy, though, when the number of studies exceeds ten. Lunny et al. [7] and Pieper et al. [6] suggest two "tools" for analyzing overlaps that scale better: (1) a citation matrix that cross-tabulates primary and secondary studies and (b) the corrected covered area (CCA). The CCA computes a single number that indicates the total overlap of primary studies for all included secondary studies. The CCA can, however, also be computed for all pairs of secondary studies in a tertiary study to give a more fine-grained overview of the overlap of primary studies. A tabular overview of the CAAs for all pairs of secondary studies results in a compact and scalable presentation of the overlap of primary studies. An example of such a CCA-matrix is shown in Fig. 3 for our example in Section 3. Bougioukas et al. [24] discuss the advantages and disadvantages of seven approaches for visualizing overlaps, including the ones shown in Figs. 1 and 3.

In our sample, two tertiary studies analyzed the overlap of primary studies between secondary studies (T15, T43), of which one (T43) used this analysis to single out the primary studies that were unique (i.e., non-overlapping) and used only those for answering its research questions. A third study (T04) went directly to the primary studies without first analyzing overlaps.

A decision about a mitigation strategy for overlaps of primary studies might not only depend on the degree of overlap and the research questions, it might also depend on the quality of the underlying secondary studies [21]. In our sample, 33 of the 47 tertiary studies reported a quality assessment of the included secondary studies. Quality assessment in tertiary studies has been investigated in detail by Costal et al. [22].

Regarding dependencies between tertiary studies, we found ten tertiary studies with dependencies.[9] The eight dependent studies listed in Table 4 (first row) plus two that these eight depend on, directly or indirectly (T35, T47). Two of the eight dependent studies (T44, T46) are extensions of T47, i.e., all three share research questions. In all three studies, double-counting was neither acknowledged as a threat for tertiary studies, in general, nor as a threat for the conducted study. In T46, it is clearly described that only secondary studies not included in T47 were considered. In T46, this can be deducted from the list of included secondary studies. The dependencies between the studies will, therefore, not lead to double-counting issues.

Of the eight dependent studies, one study (T37) builds on the search/selection results from T35, which did neither acknowledge double-counting as a threat for tertiary studies, in general nor as a threat for itself. Since there is no overlap in research questions between T37 and T35, their dependency will not lead to double-counting issues. The same can be said about T45, which reuses the search/selection results from T46 and T47.

Four tertiary studies (T05, T09, T23, T48) partially build on each other and used the same or largely overlapping sets of secondary studies, including those from T44, T46, and T47. We consider three of those (T09, T23, T48) to have double-counting bias. An analysis of their dependencies is, therefore, superfluous. T05 uses a specific subset of the search/selection results and has no overlapping research questions with any of the three other tertiary studies. We do, therefore, not consider that T05's dependence on those three leads to double-counting bias.

In our sample, we could not find any cases where double-counting issues in a tertiary study propagated to a dependent tertiary study.

[9] Actually two more but one dependency lead to the exclusion of T30, see Section 5.

8. Threats to validity

Coverage of tertiary studies in software engineering and generalizability of the findings

We used Costal et al.'s dataset [23] which is based on automated searches in Scopus and snowballing [17]. The dataset covers tertiary studies published from 2004 until early 2021. To investigate the threat of missed tertiary studies, Costal et al. [22] "conducted equivalent searches in Scopus, IEEE Xplore, ACM DL, SpringerLink, ScienceDirect, and WoS on April 28th, 2021" and concluded that no additional relevant papers were found. We are, therefore, confident that the sample of tertiary studies we have analyzed for this paper is a good sample. We did not update their results by searching for any recent tertiary studies. However, we excluded three tertiary studies from their dataset as described at the end of Section 5. Since we did not find any methodological guidelines or discussions about the double-counting issue or study overlaps in the software engineering literature, we consider this as a negligible risk. A limitation, however, is that tertiary studies published more recently might be more aware of double-counting issues and perhaps proposed additional actions to mitigate this threat.

Data extraction and analysis

As described in Section 5, we piloted the data extraction form to develop a consensus regarding what information to extract. After the data extraction phase, the extracted data for all included studies have been validated by a second co-author. To avoid conflicts of interest regarding included tertiary studies co-authored by one or more co-authors of the present study, the data extraction for this study (T50) was validated by an independent person (who is not a co-author).

Double-counting

Costal et al. deleted duplicates as well as publications "superseded by a later version from the same authors" (exclusion criterion EC1 [17, 22]). During our data extraction, we found that T11 superseded T30 and excluded T30 (i.e., we assessed T30 as a redundant tertiary study according to our terminology in Table 1).

After excluding T30, Costal et al.'s dataset contains eight tertiary studies that depend on other tertiary studies. Therefore, there is a risk of overlaps of secondary studies and that such overlaps might have propagated from one tertiary study in our dataset to a dependent one that is also included in our dataset. However, we have only investigated the tertiary studies' awareness of and handling of double-counting as well as their potential vulnerability for double-counting issues, not whether they actually did double-count. Therefore, we do not consider overlaps of secondary studies or primary studies a threat to our tertiary study.

9. Recommendations for tertiary studies

Based on the problems and mitigation strategies observed in the reviewed 47 tertiary studies, we suggest a four-step process (see Section 9.1–9.4) for dealing with the double-counting threat in tertiary studies (see Table 1 for an overview of causes of double-counting in tertiary studies). We recommend that such a process be part of the a priori design, i.e., the protocol of a tertiary study. The following data from secondary studies are required to make an informed decision about the double-counting threat in a tertiary study:

- A list of all included secondary studies.
- The research questions of the secondary studies and their data synthesis approaches.
- A list of primary studies for each of the included secondary studies to assess the overlap in primary studies.
- Once the redundancy is removed from the list of primary studies, a list of data sources, systems, cases, and populations used by the remaining unique primary studies is required to identify potential overlap of primary data in the tertiary study. Since the primary studies may have been included in different secondary studies, we cannot expect an individual secondary study to have resolved an overlap of primary data.

The data described above is needed in the four steps described below.

9.1. Step 1 – Remove duplicate and redundant secondary studies

Identify and remove any duplicates of the same secondary study. From the redundant secondary studies, use the most recent and complete version of the publication. Please see Table 1 for examples of how to identify duplicate and redundant secondary studies, respectively.

9.2. Step 2 – Judge if the overlap of primary studies is a potential threat to validity

Review the research questions and the analysis performed in a tertiary study to judge if the overlap of primary studies is a potential validity threat. This assessment needs to be made on a case-to-case basis. As a general rule of thumb, we can broadly divide tertiary studies into two categories depending on the type of information they consider (a) information about the secondary studies per se or (b) information that the secondary studies derived from primary studies.

For tertiary studies in the former category, an overlap of primary studies is no threat. Examples of such tertiary studies include studies about methodological aspects of secondary studies (e.g., about search or selection strategies in secondary studies) and studies cataloging secondary studies on a topic that only list aspects of the secondary studies (e.g., aims and scope of secondary studies, number of selected primary studies and coverage).

An example of the latter category is a tertiary study aggregating evidence regarding the effectiveness of test-driven development by using vote counting as discussed in Section 3. Double-counting due to an overlap in primary studies is a validity threat for such studies.

9.3. Step 3 – Quantify the overlap of primary studies

Map the overlap of primary studies between included secondary studies. This is done by identifying duplicate as well as redundant primary studies. Duplicates can primarily be identified automatically with tool support. The second type will require manual analysis of the titles, abstracts, and authors to identify a primary study that is redundant to one that is already included in another secondary study. From a set of redundant primary studies, the most recent and complete version should be used.

Once the duplicate and redundant primary studies have been removed, we suggest using the corrected covered area (CCA) to quantify the potential impact of the overlap of primary studies, see Section 7 for details. Use the CCA percentage range as shown in Fig. 3 as indicators for the extent of overlap between studies.

Furthermore, to identify the overlap of primary data, we should analyze the extent to which primary studies have used the same data sources, systems, cases, and populations in their investigations.

9.4. Step 4 – Address and mitigate the double-counting

For tertiary studies with a slight overlap of primary studies between their included secondary studies, we suggest that the researchers at least discuss the overlap of primary studies as a potential limitation of their study and discuss its potential impact.

For tertiary studies with a moderate or higher overlap of primary studies between their included secondary studies, we suggest that the researchers should attempt to mitigate the impact of double-counting. For example, by assessing the potential bias, the overlap might cause. Such an impact analysis is a non-trivial task and requires considering both the extent of the overlap and the quality of the primary studies shared between secondary studies.

Mitigating the threat of double-counting in software engineering (as, e.g., by Rafique and Mišić [25]) will often require re-analyzing the unique primary studies (i.e., after deleting duplicate/redundant primary studies) to answer the questions of interest for the tertiary study.

9.5. Implications for the reporting of secondary studies

This study has made the importance of specific reporting prerequisites [26] explicit for secondary studies. The following information about secondary studies is necessary to assess the extent and impact of double-counting on the results of a tertiary study:

- A clear description of related and similar secondary studies. When there are similar secondary studies, we suggest describing the overlap between the primary studies in these secondary studies using the corrected covered area (CCA, see Fig. 3 in Section 7) to facilitate the identification of (potentially) redundant secondary studies.
- An easily accessible list of primary studies included in a secondary study (preferably in a machine-readable format like BibTeX, RIS, etc.) to facilitate an analysis of overlaps of primary studies in tertiary studies that include the secondary study.
- An easily accessible list of quality scores for each of the primary studies included in a secondary study to facilitate decisions about suitable mitigation strategies in tertiary studies regarding double-counting.
- An easily accessible list of data sources, systems, cases, and populations used by the primary studies included in a secondary study to enable an assessment of potential overlap of primary data in tertiary studies using the secondary study.

10. Demonstrating the applicability of our recommendations

In this section, we evaluate to which degree our recommendations helped or would have helped to mitigate double-counting threats in our sample tertiary studies.

10.1. Handling duplicate and redundant secondary studies (step 1)

As discussed in Section 6.2, 34 of the 47 tertiary studies in our sample described that they dealt with duplicates, but only 17 of 47 explicitly noted that they deleted duplicate and redundant secondary studies. This means that for 30 of the 47 tertiary studies in our sample it is unclear whether there is a potential threat to validity due to redundant secondary studies and for 13 tertiary studies it is unclear whether a potential threat to validity due to duplicate and redundant secondary studies.

Following and documenting step 1 of our recommendations would have mitigated this issue.

10.2. Handling overlaps of primary studies (steps 2–4)

Our data shows that double-counting beyond duplicate/redundant secondary studies is a threat to validity in 19 of the 47 tertiary studies in our sample; see item #9 in Table 4.

As discussed in Section 6.4, 7 of the 47 tertiary studies in our sample provide or suggest strategies for addressing double-counting (T04, T06, T09, T10, T15, T42, T43). Six of those seven follow step 3 of our recommendation and explicitly discuss/quantify the overlap of primary studies. The seventh tertiary study (T04) goes directly to the primary studies included in the secondary studies to avoid double-counting without first analyzing overlaps, i.e., it directly jumps to step 4 of our recommendations without first quantifying the overlap (as suggested in step 3).

In Table D.8 in Appendix D, we discuss all 19 tertiary studies where double-counting beyond duplicate/redundant secondary studies is a threat to validity. For this discussion, we used the data extracted for items #11, #12, and #15. The table shows that the majority of the 19 tertiary studies would have benefited from following our recommendations. Only five of the 19 tertiary studies follow our recommendations to a large extent (T04, T10, T15, T42, T43), including the three we assessed to mitigate double-counting threats beyond duplicate/redundant secondary studies sufficiently (T04, T10, T43; see item #10 in Table 4).

Taken together, we can say that our recommendations would have helped to mitigate double-counting threats in many cases or at least made it explicit for readers that double-counting has been considered and sufficiently addressed in the study.

11. Summary and conclusions

We discussed issues concerning double-counting in tertiary studies and exemplified in which ways double-counting may affect research quality. We furthermore analyzed 47 tertiary studies in software engineering and found that double-counting is an overlooked issue in those. For tertiary studies focusing on information about primary research, double-counting is a potential threat to validity. We, therefore, recommend documenting and analyzing the overlap of primary studies and suggest tools borrowed from the health sciences to do so (see Section 7, specifically Fig. 3). Furthermore, we recommend examining the threats to validity that these overlaps may cause and reporting how they were addressed or mitigated.

We also proposed recommendations for dealing with the double-counting threat in tertiary studies. An application of the recommendations on the 47 tertiary studies in our sample showed promising results. The recommendations would have helped the tertiary studies' authors identify, assess, and choose mitigation strategies to deal with the threat of double-counting.

Declaration of competing interest

No author associated with this paper has disclosed any potential or pertinent conflicts which may be perceived to have impending conflict with this work. For full disclosure statements refer to https://doi.org/10.1016/j.infsof.2023.107174.

Data availability

The data is shared in an online supplement, see Section 6.

Acknowledgments

This work has been supported by ELLIIT, the Strategic Research Area within IT and Mobile Communications, funded by the Swedish Government.

Appendix A. List of shared primary studies in Fig. 1

Table A.5 shows the secondary and primary studies that were the subject of the discussion of overlaps in Section 3. The paper IDs for the secondary studies map those in Table A.5.

Table A.5
Shared primary studies (P1–P6) in the four secondary studies (S02, S07, S10, S11) in Nurdiani et al. [4].

ID	Reference
S02	A. Causevic, D. Sundmark, and S. Punnekkat. Factors limiting industrial adoption of test driven development: A systematic review. Proceedings of the Fourth IEEE International Conference on Software Testing, Verification and Validation, pages 337–346, 2011
S07	H. Munir, M. Moayyed, and K. Petersen. Considering rigor and relevance when evaluating test driven development: A systematic review. Information and Software Technology, 56(4):375–394, 2014
S10	P. Sfetsos and I. Stamelos. Empirical studies on quality in agile practices: A systematic literature review. Proceedings of the Seventh International Conference on the Quality of Information and Communications Technology, pages 44–53, 2010
S11	Y. Rafique and V. B. Mišić. The effects of test-driven development on external quality and productivity: A meta-analysis. IEEE Transactions on Software Engineering, 39(6):835–856, 2013
P1	N. Nagappan, E. M. Maximilien, T. Bhat, and L. Williams. Realizing quality improvement through test driven development: results and experiences of four industrial teams. Empirical Software Engineering, 13(3):289–302, 2008
P2	L. Williams, E. M. Maximilien, and M. Vouk. Test-driven development as a defect-reduction practice. Proceedings of the 14th International Symposium on Software Reliability Engineering, pages 34–45, 2003
P3	L. Huang and M. Holcombe. Empirical investigation towards the effectiveness of test first programming. Information and Software Technology, 51(1):182–194, 2009
P4	H. Erdogmus, M. Morisio, and M. Torchiano. On the effectiveness of the test-first approach to programming. IEEE Transactions on software Engineering, 31(3):226–237, 2005
P5	M. M. Mueller and O. Hagner. Experiment about test-first programming. IEE Proceedings-Software, 149(5):131–136, 2002
P6	A. Gupta and P. Jalote. An experimental evaluation of the effectiveness and efficiency of the test driven development. Proceedings of the First International Symposium on Empirical Software Engineering and Measurement, pages 285–294, 2007

Appendix B. Mapping between RQs and data extraction items

See Table B.6.

Table B.6
Mapping of research questions to data extraction items in Table 3.

		Data extraction items						
		#7	#8	#9	#10	#11	#12	#13
Research questions	RQ1 – mindful	x	x				x	x
	RQ2 – double counting			x	x	x	x	x
	RQ3 – mitigated			x	x	x		x
	RQ4 – mitigation approach						x	x

Appendix C. List of tertiary studies

Table C.7 below lists the 50 tertiary studies in software engineering (T01–T50) originally selected by Costal et al. [17]. Of those 50, we excluded three tertiary studies, T08, T30, and T39.

T08 was excluded since it is a hybrid secondary/tertiary study. In its abstract, it states that the authors used primary studies on the topic of interest.

T30 was excluded since it is redundant to T11.

T39 was excluded since it is a hybrid secondary/tertiary study. T39's title and search string indicate that it is a tertiary study. However, its inclusion criteria indicate that being a secondary study was no requirement for inclusion. To support our decision, we obtained the list of studies included in T39. Since this list contains primary studies, we excluded T39 from our dataset.

Table C.7
List of the 50 tertiary studies selected by Costal et al. [17]. The tertiary studies excluded for the present study are shown with striked-through IDs (~~T08~~, ~~T30~~, ~~T39~~).

ID	Reference
T01	H. Cadavid, V. Andrikopoulos, and P. Avgeriou. Architecting systems of systems: A tertiary study. Information and Software Technology, 118:106202, 2020
T02	J. L. Barros-Justo, F. B. Benitti, and S. Matalonga. Trends in software reuse research: A tertiary study. Computer Standards & Interfaces, 66:103352, 2019
T03	M. U. Khan, S. Sherin, M. Z. Iqbal, and R. Zahid. Landscaping systematic mapping studies in software engineering: A tertiary study. Journal of Systems and Software, 149:396–436, 2019
T04	N. Rios, M. G. de Mendonça Neto, and R. O. Spínola. A tertiary study on technical debt: Types, management strategies, research trends, and base information for practitioners. Information and Software Technology, 102:117–145, 2018
T05	D. Budgen, P. Brereton, S. Drummond, and N. Williams. Reporting systematic reviews: Some lessons from a tertiary study. Information and Software Technology, 95:62–74, 2018
T06	R. Hoda, N. Salleh, J. Grundy, and H. M. Tee. Systematic literature reviews in agile software development: A tertiary study. Information and software technology, 85:60–70, 2017
T07	V. Garousi and M. V. M̀antyl´a. A systematic literature review of literature reviews in software testing. Information and Software Technology, 80:195–216, 2016
~~T08~~	Y. Shakeel, J. Krüger, I. V. Nostitz-Wallwitz, G. Saake, and T. Leich. Automated selection and quality assessment of primary studies: A systematic literature review. Journal of Data and Information Quality, 12(1):1–26, 2019
T09	D. Budgen, P. Brereton, N. Williams, and S. Drummond. The contribution that empirical studies performed in industry make to the findings of systematic reviews: A tertiary study. Information and software technology, 94:234–244, 2018
T10	T. N. Kudo, R. F. Bulcão-Neto, and A. M. Vincenzi. Requirement patterns: A tertiary study and a research agenda. IET Software, 14(1):18–26, 2020
T11	L. Yang, H. Zhang, H. Shen, X. Huang, X. Zhou, G. Rong, and D. Shao. Quality assessment in systematic literature reviews: A software engineering perspective. Information and Software Technology, 130:106397, 2021
T12	K. Curcio, R. Santana, S. Reinehr, and A. Malucelli. Usability in agile software development: A tertiary study. Computer Standards & Interfaces, 64:61–77, 2019
T13	M. Goulão, V. Amaral, and M. Mernik. Quality in model-driven engineering: A tertiary study. Software Quality Journal, 24(3):601–633, 2016
T14	M. Raatikainen, J. Tiihonen, and T. M̀annist´o. Software product lines and variability modeling: A tertiary study. Journal of Systems and Software, 149:485–510, 2019
T15	I. Nurdiani, J. Börstler, and S. A. Fricker. The impacts of agile and lean practices on project constraints: A tertiary study. Journal of Systems and Software, 119:162–183, 2016
T16	G. T. G. Neto, W. B. Santos, P. T. Endo, and R. A. Fagundes. Multivocal literature reviews in software engineering: Preliminary findings from a tertiary study. Proceedings of the 13th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, pp 1–6, 2019
T17	A. Idri and L. Cheikhi. A survey of secondary studies in software process improvement. Proceedings of the 13th ACS/IEEE International Conference of Computer Systems and Applications, pp 1–8, 2016
T18	X. Zhou, Y. Jin, H. Zhang, S. Li, and X. Huang. A map of threats to validity of systematic literature reviews in software engineering. Proceedings of the 23rd Asia-Pacific Software Engineering Conference, pp 153–160, 2016
T19	S. P. Pillai, S. Madhukumar, and T. Radharamanan. Consolidating evidence based studies in software cost/effort estimation – a tertiary study. Proceedings of the 2017 IEEE Region 10 Conference, pp 833–838, 2017
T20	A. Yasin, R. Fatima, L. Wen, W. Afzal, M. Azhar, and R. Torkar. On using grey literature and google scholar in systematic literature reviews in software engineering. IEEE Access, 8:36226–36243, 2020
T21	J. Krüger, C. Lausberger, I. von Nostitz-Wallwitz, G. Saake, and T. Leich. Search. review. repeat? an empirical study of threats to replicating slr searches. Empirical Software Engineering, 25(1):627–677, 2020
T22	E. Bayram, B. Doğan, and V. Tunalı. Bibliometric analysis of the tertiary study on agile software development using social network analysis. Proceedings of the Innovations in Intelligent Systems and Applications Conference, pp 1–4, 2020
T23	D. Budgen, P. Brereton, N. Williams, and S. Drummond. What support do systematic reviews provide for evidence – informed teaching about software engineering practice? e-Informatica Software Engineering Journal, 14(1):7–60, 2020
T24	V. Delavari, E. Shaban, M. Janssen, and A. Hassanzadeh. Thematic mapping of cloud computing based on a systematic review: A tertiary study. Journal of Enterprise Information Management, 33(1):161–190, 2020
T25	G. A. García-Mireles and M. E. Morales-Trujillo. Gamification in software engineering: A tertiary study. Proceedings of the 8th International Conference on Software Process Improvement, pp 116–128, 2019
T26	P. A. Duarte, F. M. Barreto, P. A. Aguilar, J. Boudy, R. M. Andrade, and W. Viana. Aal platforms challenges in iot era: A tertiary study. Proceedings of the 13th Annual Conference on System of Systems Engineering, pp 106–113, 2018
T27	C. Fu, H. Zhang, X. Huang, X. Zhou, and Z. Li. A review of meta-ethnographies in software engineering. Proceedings of the Evaluation and Assessment on Software Engineering, pp 68–77, 2019
T28	B. Napoleão, K. R. Felizardo, É. F. de Souza, and N. L. Vijaykumar. Practical similarities and differences between systematic literature reviews and systematic mappings: a tertiary study. Proceedings of the 29th International Conference on Software Engineering and Knowledge Engineering, pp 85–90, 2017

(*continued on next page*)

Table C.7 (*continued*).

T29	P. Singh, M. Galster, and K. Singh. How do secondary studies in software engineering report automated searches? A preliminary analysis. Proceedings of the 22nd International Conference on Evaluation and Assessment in Software Engineering, pp 145–150, 2018
~~T30~~	Y. Zhou, H. Zhang, X. Huang, S. Yang, M. A. Babar, and H. Tang. Quality assessment of systematic reviews in software engineering: A tertiary study. Proceedings of the 19th international conference on evaluation and assessment in software engineering, pp 1–14, 2015
T31	L. Villalobos Arias, C. U. Quesada López, A. Martínez Porras, and M. Jenkins Coronas. A tertiary study on model-based testing areas, tools and challenges: Preliminary results. Proceedings of the 21st Iberoamerican Conference on Software Engineering, pp 15–28, 2018
T32	A. Ampatzoglou, S. Bibi, P. Avgeriou, M. Verbeek, and A. Chatzigeorgiou. Identifying, categorizing and mitigating threats to validity in software engineering secondary studies. Information and Software Technology, 106:201–230, 2019
T33	A. A. Khan, J. Keung, M. Niazi, S. Hussain, and H. Zhang. Systematic literature reviews of software process improvement: A tertiary study. Proceedings of the 24th European Conference on Software Process Improvement, pp 177–190, 2017
T34	C. Marimuthu and K. Chandrasekaran. Systematic studies in software product lines: A tertiary study. Proceedings of the 21st International Systems and Software Product Line Conference–Volume A, pp 143–152, 2017
T35	H. Zhang and M. A. Babar. Systematic reviews in software engineering: An empirical investigation. Information and software technology, 55(7):1341–1354, 2013
T36	D. S. Cruzes and T. Dybå. Research synthesis in software engineering: A tertiary study. Information and Software Technology, 53(5):440–455, 2011
T37	H. Tang, Y. Zhou, X. Huang, and G. Rong. Does Pareto's law apply to evidence distribution in software engineering? An initial report. Proceedings of the Third International Workshop on Evidential Assessment of Software Technologies, pp 9–16, 2014
T38	M. Bano, D. Zowghi, and N. Ikram. Systematic reviews in requirements engineering: A tertiary study. Proceedings of the 4th IEEE International Workshop on Empirical Requirements Engineering, pp 9–16, 2014
~~T39~~	S. Imtiaz, M. Bano, N. Ikram, and M. Niazi. A tertiary study: Experiences of conducting systematic literature reviews in software engineering. Proceedings of the 17th International Conference on Evaluation and Assessment in Software Engineering, pp 177–182, 2013
T40	N. Salleh and A. Nordin. Trends and perceptions of evidence-based software engineering research in Malaysia. Proceedings of the 5th International Conference on Information and Communication Technology for The Muslim World, pp 1–6, 2014
T41	A. B. Marques, R. Rodrigues, and T. Conte. Systematic literature reviews in distributed software development: A tertiary study. Proceedings of the Seventh IEEE International Conference on Global Software Engineering, pp 134–143, 2012
T42	J. M. Verner, O. P. Brereton, B. A. Kitchenham, M. Turner, and M. Niazi. Risks and risk mitigation in global software development: A tertiary study. Information and Software Technology, 56(1):54–78, 2014
T43	G. K. Hanssen, D. Šmite, and N. B. Moe. Signs of agile trends in global software engineering research: A tertiary study. Proceedings of the Sixth IEEE International Conference on Global Software Engineering Workshop, pp 17–23, 2011
T44	F. Q. Da Silva, A. L. Santos, S. Soares, A. C. C. França, C. V. Monteiro, and F. F. Maciel. Six years of systematic literature reviews in software engineering: An updated tertiary study. Information and Software Technology, 53(9):899–913, 2011
T45	F. Q. Da Silva, A. L. Santos, S. C. Soares, A. C. C. França, and C. V. Monteiro. A critical appraisal of systematic reviews in software engineering from the perspective of the research questions asked in the reviews. Proceedings of the Fourth International Symposium on Empirical Software Engineering and Measurement, pp 1–4, 2010
T46	B. Kitchenham, R. Pretorius, D. Budgen, O. P. Brereton, M. Turner, M. Niazi, and S. Linkman. Systematic literature reviews in software engineering – a tertiary study. Information and software technology, 52(8):792–805, 2010
T47	B. Kitchenham, O. P. Brereton, D. Budgen, M. Turner, J. Bailey, and S. Linkman. Systematic literature reviews in software engineering – A systematic literature review. Information and software technology, 51(1):7–15, 2009
T48	D. Budgen, S. Drummond, P. Brereton, and N. Holland. What scope is there for adopting evidence – Informed teaching in software engineering? In . Proceedings of the 34th International Conference on Software Engineering, pp 1205–1214, 2012
T49	X. Huang, H. Zhang, X. Zhou, M. A. Babar, and S. Yang. Synthesizing qualitative research in software engineering: A critical review. Proceedings of the 40th International Conference on Software Engineering, pp 1207–1218, 2018
T50	K. Petersen and N. B. Ali. Identifying strategies for study selection in systematic reviews and maps. Proceedings of the International Symposium on Empirical Software Engineering and Measurement, pp 351–354, 2011

Appendix D. Discussion of tertiary studies referred to in Section 10.2

See Table D.8.

Table D.8
Review of the 19 tertiary studies where double-counting beyond duplicate/redundant secondary studies is a threat to validity for the study according to our assessment (see item #9 in Table 4).

ID	Discussion wrt to our recommendations in Sections 9.2–9.4
T01	In Figures 9–11 of T01, results from the included secondary studies are aggregated. The aggregated data might be affected by overlapping primary studies. T01 noted that the extent of overlapping primary studies could not be assessed since the primary studies included in some of the secondary studies could not be identified (see footnote 10 in T01). A partial quantification of the overlap of primary studies (step 3) would have provided a clearer picture of the potential extent of the problem.
T02	T02 covers 56 secondary studies with a total of 2640 primary studies. The aggregated data in Figure 10 is prone to double-counting due to overlapping primary studies. Assessing the potential overlap of primary studies (steps 2–4) would have helped to mitigate the threat.
T04	T04 successfully mitigated the threat by discarding duplicate primary studies (see Appendix C in T04). Deleting duplicate/redundant primary studies is a good strategy to mitigate the threat (see step 4 of our recommendations).
T06	T06 did not follow steps 2–4 of our recommendations and explains that "an analysis of the overlap between the sets of primary studies was not performed. This has particular reference to RQ5 since a potential high level of overlap of primary studies between SLRs in the same research area can provide a skewed view of the progress achieved." Following our recommendations would have helped to mitigate this problem.
T09	T09 argues that the overlap of primary studies "is likely to be low". The issue is not discussed further. Quantifying the overlap, as suggested in step 3 of our recommendations would have helped to support the claim with evidence.
T10	T10 analyzed the overlap of primary studies and found that the overlap is small. Of the 50 primary studies, 44 are unique, and 4 of the unique ones are shared by two or three secondary studies. I.e., T10 followed our recommendations (steps 3–4) which helped T10 to mitigate the double-counting threat.
T12	Table 6 in T12 might be biased by overlapping primary studies. Following step 3 of our recommendations (quantifying the overlap) would have helped to clarify whether the overlap of primary studies is a potential threat to validity.
T13	Tables 6 and 7 in T13 might be biased by overlapping primary studies. Following step 3 of our recommendations (quantifying the overlap) would have helped to clarify whether the overlap of primary studies is a potential threat to validity.
T14	Tables 9–15 in T14 might be biased by overlapping primary studies. Following step 3 of our recommendations (quantifying the overlap) would have helped to clarify whether the overlap of primary studies is a potential threat to validity.
T15	Tables 8–9 and 13–26 might be biased by overlapping primary studies. T15 is the only tertiary study in our sample with a research question related to overlapping primary studies (RQ1.1). T15 also provides a partial analysis of the extent of overlap as we suggest in step 3 of our recommendations. Fig. 1 in our Section 3 gives an overview of the analysis in T15. Although the overlap in T15 is very high (as we show in our Fig. 3, the authors of T15 "believe that reviewing overlapping papers would not have changed the outcome of our study" and do not mitigate the threat. Following our recommendation of computing the CCA percentage (see step 3 and Fig. 3) would have shown that the overlap is very high and needs to be addressed.
T17	There are several aggregations of data that might be prone to double-counting of primary studies. The issue is not discussed in T17. Following steps 2–4 of our recommendations would have helped to clarify whether the overlap of primary studies is a potential threat to validity.
T19	Tables I–III are prone to double-counting of primary studies. The issue is not discussed in T19. Following steps 2–4 of our recommendations would have helped to clarify whether the overlap of primary studies is a potential threat to validity.
T23	Most findings and tables are potentially prone to double-counting of primary studies. The issue is not discussed in T23. Following steps 2–4 of our recommendations would have helped to clarify whether the overlap of primary studies is a potential threat to validity.
T24	T24 coded the secondary studies' contexts and findings to find themes. The themes were then used to calculate statistics for relationships between themes. According to our assessment, this approach is sensitive to the double-counting of primary studies. Double-counting of primary studies is not discussed in T24. Following steps 2–4 of our recommendations would have helped to clarify whether the overlap of primary studies is a potential threat to validity.
T25	When aggregating gaming elements (in Section 4.3 of T25), the potential double counting of primary studies is not considered. The percentages computed in Section 4.4 might also be affected by the double-counting of primary studies. Following steps 2–4 of our recommendations would have helped to clarify whether the overlap of primary studies is a potential threat to validity.
T38	When answering RQ3 (gaps in the coverage of RE research topics in the published SLRs), T38 compares aggregation that might be biased by overlapping primary studies. Following steps 2–4 of our recommendations would have helped to clarify whether the overlap of primary studies is a potential threat to validity.
T42	T42 discusses the issue as a limitation of the study but provides raw data (1st bullet in Sect 5): "We could not sum support provided by the different studies, as the primary papers referred to by the SLRs in many cases overlapped; hence we would have been counting the same empirical support twice. This made it difficult to be sure about the real degree of empirical support for many items; however, the tables in Section 4.6 supply the reader with the raw data so they can make up their own minds." Considering how T42 handles the potential double-counting of primary studies, we would argue that T42 almost follows our recommendations. A CCA matrix with details about the extent of the overlap, as we suggest in step 3, would have provided evidence and actionable information about the extent of the overlap.
T43	T43 collects primary studies from the selected secondary studies and identifies/rectifies the overlap of primary studies before answering the trend question. T43 also provides a graphical overview of the overlap of primary studies but does not quantify the overlap. A CCA matrix with details about the extent of the overlap, as we suggest in step 3, would have provided evidence and actionable information about the extent of the overlap.
T48	T48 did not analyze the extent of overlap of primary studies when aggregating data and claiming that "coverage of the major SEEK headings is uneven." Following steps 2–4 of our recommendations would have helped to clarify whether the overlap of primary studies is a potential threat to validity.

References

[1] B. Kitchenham, O.P. Brereton, D. Budgen, M. Turner, J. Bailey, S. Linkman, Systematic literature reviews in software engineering — A systematic literature review, Inf. Softw. Technol. 51 (1) (2009) 7–15.

[2] M.U. Khan, S. Sherin, M.Z. Iqbal, R. Zahid, Landscaping systematic mapping studies in software engineering: A tertiary study, J. Syst. Softw. 149 (2019) 396–436.

[3] B. Kitchenham, S. Charters, Guidelines for Performing Systematic Literature Reviews in Software Engineering, Report EBSE-2007-1, Keele University, 2007.

[4] I. Nurdiani, J. Börstler, S.A. Fricker, The impacts of agile and lean practices on project constraints: A tertiary study, J. Syst. Softw. 119 (2016) 162–183.

[5] H.K.V. Tran, M. Unterkalmsteiner, J. Börstler, N. bin Ali, Assessing test artifact quality – A tertiary study, Inf. Softw. Technol. 139 (2021) 106620.

[6] D. Pieper, S.-L. Antoine, T. Mathes, E.A. Neugebauer, M. Eikermann, Systematic review finds overlapping reviews were not mentioned in every other overview, J. Clin. Epidemiol. 67 (4) (2014) 368–375.

[7] C. Lunny, D. Pieper, P. Thabet, S. Kanji, Managing overlap of primary study results across systematic reviews: Practical considerations for authors of overviews of reviews, BMC Med. Res. Methodol. 21 (1) (2021) 1–14.

[8] S.J. Senn, Overstating the evidence – double counting in meta-analysis and related problems, BMC Med. Res. Methodol. 9 (1) (2009) 1–7.

[9] J.M. Verner, O.P. Brereton, B.A. Kitchenham, M. Turner, M. Niazi, Risks and risk mitigation in global software development: A tertiary study, Inf. Softw. Technol. 56 (1) (2014) 54–78.

[10] F. Alfonso, J. Bermejo, J. Segovia, Duplicate or redundant publication: Can we afford it? Rev. Esp. Cardiol. 58 (5) (2005) 601–604 (English Edition).

[11] J. Sayyad Shirabad, T. Menzies, The PROMISE Repository of Software Engineering Databases, School of Information Technology and Engineering, University of Ottawa, Canada, 2005, URL: http://promise.site.uottawa.ca/SERepository.

[12] H. Do, S. Elbaum, G. Rothermel, Supporting controlled experimentation with testing techniques: An infrastructure and its potential impact, Empir. Softw. Eng. 10 (4) (2005) 405–435, The repository is available at https://sir.csc.ncsu.edu/portal/index.php.

[13] N. bin Ali, E. Engström, M. Taromirad, M.R. Mousavi, N.M. Minhas, D. Helgesson, S. Kunze, M. Varshosaz, On the search for industry-relevant regression testing research, Empir. Softw. Eng. 24 (4) (2019) 2020–2055, http://dx.doi.org/10.1007/s10664-018-9670-1.

[14] M. Cruz, B. Bernárdez, A. Durán, J.A. Galindo, A. Ruiz-Cortés, Replication of studies in empirical software engineering: A systematic mapping study, from 2013 to 2018, IEEE Access 8 (2019) 26773–26791.

[15] H. Munir, M. Moayyed, K. Petersen, Considering rigor and relevance when evaluating test driven development: A systematic review, Inf. Softw. Technol. 56 (4) (2014) 375–394.

[16] A. Ampatzoglou, S. Bibi, P. Avgeriou, M. Verbeek, A. Chatzigeorgiou, Identifying, categorizing and mitigating threats to validity in software engineering secondary studies, Inf. Softw. Technol. 106 (2019) 201–230.

[17] D. Costal, C. Farré, X. Franch, C. Quer, Inclusion and exclusion criteria in software engineering tertiary studies: A systematic mapping and emerging framework, in: F. Lanubile, M. Kalinowski, M.T. Baldassarre (Eds.), Proceedings of the 15th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2021, pp. 30:1–30:6.

[18] B.A. Kitchenham, D. Budgen, P. Brereton, Evidence-Based Software Engineering and Systematic Reviews, CRC Press, 2016.

[19] N. Rios, M.G. de Mendonça Neto, R.O. Spínola, A tertiary study on technical debt: Types, management strategies, research trends, and base information for practitioners, Inf. Softw. Technol. 102 (2018) 117–145.

[20] M. Pollock, R.M. Fernandes, L.A. Becker, D. Pieper, L. Hartling, Chapter V: Overviews of reviews, in: J.P. Higgins, J. Thomas, J. Chandler, M. Cumpston, T. Li, M.J. Page, V.A. Welch (Eds.), Cochrane Handbook for Systematic Reviews of Interventions, Version 6.2, Cochrane, 2021, URL: https://training.cochrane.org/handbook/current/chapter-v.

[21] M. Gates, A. Gates, S. Guitard, M. Pollock, L. Hartling, Guidance for overviews of reviews continues to accumulate, but important challenges remain: a scoping review, Syst. Rev. 9 (1) (2020) 1–19.

[22] D. Costal, C. Farré, X. Franch, C. Quer, How tertiary studies perform quality assessment of secondary studies in software engineering, in: T. Conte, M. Solari, S.S. Reinehr, R. Prikladnicki, N. Condori-Fernández, V.E.S. Souza, N.M.C. Valentim, S. Martínez-Fernández, M. Genero, M. Kalinowski, M. Jenkins, A. Martínez, C. Quesada-López (Eds.), Proceedings of the XXIV Iberoamerican Conference on Software Engineering, 2021, pp. 56–69.

[23] D. Costal, C. Farré, X. Franch, C. Quer, How tertiary studies perform quality assessment of secondary studies in software engineering – Replication package, 2021, http://dx.doi.org/10.5281/zenodo.5094807.

[24] K.I. Bougioukas, E. Vounzoulaki, C.D. Mantsiou, E.D. Savvides, C. Karakosta, T. Diakonidis, A. Tsapas, A.-B. Haidich, Methods for depicting overlap in overviews of systematic reviews: An introduction to static tabular and graphical displays, J. Clin. Epidemiol. 132 (2021) 34–45.

[25] Y. Rafique, V.B. Mišić, The effects of test-driven development on external quality and productivity: A meta-analysis, Trans. Softw. Eng. 39 (6) (2012) 835–856.

[26] M. Usman, N. bin Ali, C. Wohlin, A quality assessment instrument for systematic literature reviews in software engineering, CoRR abs/2109.10134, 2021, arXiv:2109.10134.