# Cohort Studies in Software Engineering: A Vision of the Future

Nyyti Saarimäki
Tampere University
Tampere, Finland
nyyti.saarimaki@tuni.fi

Valentina Lenarduzzi
LUT University
Lahti, Finland
valentina.lenarduzzi@lut.fi

Sira Vegas
Universidad Politécnica de Madrid
Madrid, Spain
svegas@fi.upm.es

Natalia Juristo
Universidad Politécnica de Madrid
Madrid, Spain
natalia@fi.upm.es

Davide Taibi
Tampere University
Tampere, Finland
davide.taibi@tuni.fi

## ABSTRACT

*Background.* Most Mining Software Repositories (MSR) studies cannot obtain causal relations because they are not controlled experiments. The use of cohort studies as defined in epidemiology could help to overcome this shortcoming.
*Objective.* Propose the adoption of cohort studies in MSR research in particular and empirical Software Engineering (SE) in general.
*Method.* We run a preliminary literature review to show the current state of the practice of cohort studies in SE. We explore how cohort studies overcome the issues that prevent the identification of causality in this type of non-experimental designs.
*Results.* The basic mechanism used by cohort studies to try to obtain causality consists of controlling potentially confounding variables. This is articulated by means of different techniques.
*Conclusion.* Cohort studies seem to be a promising approach to be used in MSR in particular and SE in general.

## CCS CONCEPTS

• **Software and its engineering**;

## KEYWORDS

Cohort Study, Empirical Software Engineering, Empirical Methods

**ACM Reference Format:**
Nyyti Saarimäki, Valentina Lenarduzzi, Sira Vegas, Natalia Juristo, and Davide Taibi. 2020. Cohort Studies in Software Engineering: A Vision of the Future. In *ACM / IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM) (ESEM '20), October 8–9, 2020, Bari, Italy.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3382494.3422160

## 1 INTRODUCTION

Studies from the Mining Software Repositories (MSR) field commonly claim that the observed relationships between the examined variables cannot be proven to be causal [1, 18, 28]. The reason being that they perform correlational studies, and therefore, cannot declare causation [21]. The main reason for not being able to run controlled experiments in MSR is that it is not possible to randomly allocate experimental units to treatments or to manipulate the treatments to be studied as the data set typically exists (has already been collected) when the investigation starts.

Juristo [15] performed a small-scale Systematic Mapping Study (SMS) to get evidence on the use of the term *experiment* in the MSR literature. Out of the 46 obtained primary studies, the author concludes that 22 are not controlled experiments and further checks are needed for the remaining 24. She suggests that these 24 studies are compatible with the definition of *observational study* given in epidemiology.

Epidemiologists use *cohort studies*—a type of observational study—when they are searching for cause-effect relationships and it is not possible to run controlled experiments. A well-known success story of their use in epidemiology is the identification of smoking as a cause for the development of lung cancer[1] [4].

Note that the situation in which epidemiologists use cohort studies is similar to the one that takes place in the MSR field. In MSR, researchers are also interested in identifying causality, but they cannot perform experiments either. As an example, several studies in the past have investigated the impact of cyclomatic complexity on change-proneness [1, 18, 28]. These studies concluded that there is a high probability that increased cyclomatic complexity might lead to an increase of change-proneness in software systems, but no causal relationship could be identified.

In this work, we propose that the MSR field incorporates the use of cohort studies as defined in epidemiology to increase the level of evidence obtained by their studies. This should help the MSR field mature. To achieve this goal, the concept of how to run cohort studies in MSR needs to be understood and adapted from epidemiology. This type of task has been successfully performed in the past in Software Engineering (SE) with other types of studies, like controlled (quasi-)experiments, case studies, and Systematic Literature Reviews (SLRs). The 10 year goal is adding cohort studies to the MSR—and therefore to the Empirical Software Engineering (ESE)—toolbox. In order to address this goal, in this paper we set the foundations on which future research on this topic can be developed.

The paper has been structured as follows: Section 2 provides background knowledge about observational and cohort studies;

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
*ESEM '20, October 8–9, 2020, Bari, Italy*
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-7580-1/20/10…$15.00
https://doi.org/10.1145/3382494.3422160

[1] In this case, it would be unethical to run an experiment, as it would imply asking a large set of non-smokers to start smoking.

Section 3 illustrates, by means of an example, a possible future where cohort studies are common practice in MSR/SE; Section 4 analyzes the current state of the practice of cohort studies in SE; Section 5 sets the foundations for a future definition of cohort studies in SE; Section 6 shows our proposed roadmap; and Section 7 concludes.

## 2 BACKGROUND

### 2.1 Observational Studies

Sometimes it is not possible to conduct a controlled experiment. This could be due to ethical reasons (i.e. exposition to something known or suspected to be harmful), rare dependent variable, or significant amount of resources required to conduct the study.

In epidemiology, the alternative to controlled experiments are *observational studies*. Even though observational studies are not considered to provide as high level of evidence as controlled experiments, they are considered the next best study methodology [27].

Figure 1 shows the classification of studies in epidemiological research [11]. We can see that observational studies are non-interventional, i.e., researchers do not control the level of exposure to factors in study participants but instead they only observe and collect data on a selected population. Additionally, we can see that there are two main types of observational studies: those that use comparison groups (analytical studies), and those that lack it, and only record the current state (descriptive studies).

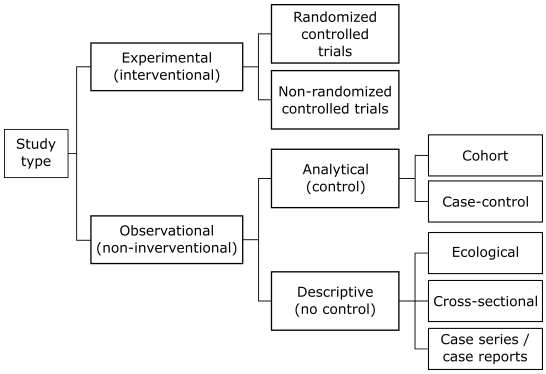


**Figure 1: Types of observational studies.**

In this research we focus on cohort studies. The reasons for this are that cohort studies are:

- The closest to controlled experiments, following experimental studies in the pyramid of evidence (see Figure 2 [9]).
- Analytical studies. They aim at understanding how exposure to different treatments affects the value of a dependent variable.

Cohort studies are well understood in epidemiology as they are one of the most adopted observational study types [10, 12].

### 2.2 Cohort Studies

This section outlines cohort studies as they are done in epidemiology. In a cohort study two (or more) groups of individuals are selected from a population of interest [10, 12]. One group (or more)

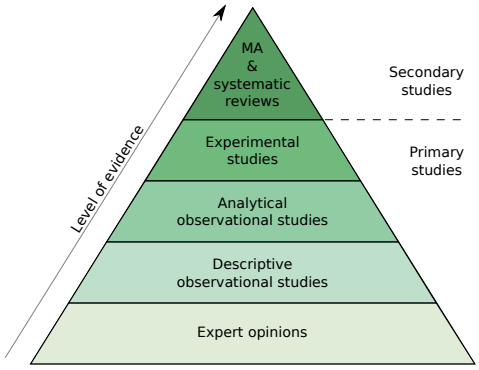


**Figure 2: Pyramid of evidence.**

consists of subjects exposed to a selected factor of interest (or different levels of the factor), while the subjects in the other group are not exposed at all. The exposure is measured for each individual at the beginning of the study (baseline). The subjects are then followed for a specified period of time—that varies depending on the research questions, available data, and resources—in order to determine whether they develop the outcome. The gathered data is then used to analyze whether there is an association between the exposure and the outcome. Figure 3 shows how data is collected in a cohort study [27].

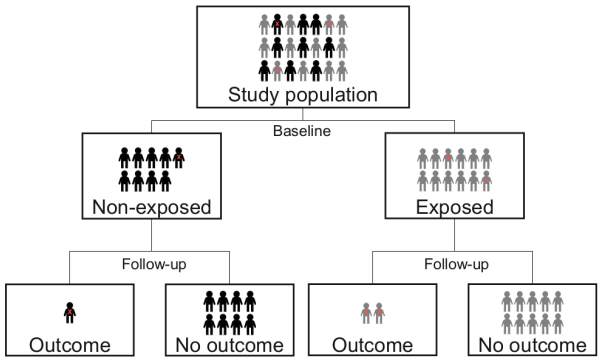


**Figure 3: Data collection in a cohort study.**

A cohort study can be done prospectively or retrospectively. If the study is done from the present time to the future, it is a *prospective* study. In this case the data about the exposures and outcomes is collected during the follow-up. In *retrospective* studies the researchers examine data that already exist, because it has been collected in the past.

Regardless of when the data is collected, it is always analyzed from exposure to outcome. The researchers select a start date for the study and initiate the data collection from that date on. Note that in retrospective studies the start date is in the past. But even in this case, the selection of individuals is made based on the data previous to the start date. Retrospective studies require less resources than prospective studies, and can be done faster—as they use existing data. However, the downside is that some needed variable might not have been measured.

A major strength in cohort studies is that the temporal relationship between the exposure and the outcome is clear, making it possible to understand causality.

## 3 A POSSIBLE FUTURE

In this section, we describe a fictional future, based on a real research question that is being investigated today in the context of MSR.

Previous research indicates that classes affected by the code smell *Long Method* (LM) are more fault-prone than other classes [23]. However, it is not clear whether the observed association is an effect caused by LM or it is due to other reasons, such as the class size or the number of people developing the class [23].

A researcher decides to further investigate this research question. Note that a controlled experiment in this setting—professionals working with real software—would not be feasible due to the amount of resources needed. The context chosen is a set of Java open-source projects from Apache Software Foundation, with projects of a given size, age and complexity, following the guidelines provided by Nagappan et al. [22].

The research follows existing guidelines about how to conduct cohort studies in SE. This means that some issues need to be taken into consideration differently from controlled experiments:

- The existence of other variables potentially affecting results must be controlled during the design of the study. In this specific case, two variables are identified: class size, and number of people developing each class. Therefore, class selection is made so that the number of classes with LM and without LM are balanced in terms of class size and number of developers coding the class.
- Data about fault-proneness is collected.
- The possibility that other variables are affecting results must be explicitly ruled out during analysis. This involves running analyses that adjust for the effect of class size and number of people developing each class. Finally, a sensitivity analysis is run to rule out the influence of other possible unknown variables.

At the end of the study, the researcher has ruled out the effect of known and unknown variables affecting the results. After the study is repeated several times, consistent results are obtained, and other possible variables affecting the results are identified and ruled out, confidence is increased in the existence of causality.

## 4 STATE OF THE PRACTICE

In this Section we present the results of a preliminary review of the current use of the cohort studies in ESE literature. In Section 4.1 we discuss how the key concepts are currently understood in ESE and present studies using methodologies resembling to cohort study methodology. In Section 4.2 we present the guidelines for the different study types currently used in SE, and finally in Section 4.3 we discuss the guidelines applied in current MSR studies.

### 4.1 Cohort Studies in Empirical Software Engineering

This section gives a preliminary view on how cohort studies are currently present in the SE literature. We conducted a preliminary literature review in Scopus and Google Scholar by combining the key word "software engineering" with "cohort", "cohort study", "field study", and "observational study".

We noticed that the term *cohort* in ESE is typically used to refer to the group of study subjects—a group of subjects that are similar to each other regarding some variable(s), such as age, size, or experience [14]. In addition, the term observational seems to be often used as a synonym for an observational case study. Therefore, in a large portion of SE studies that mention the term "cohort studies", they do not refer to the study type used in epidemiology.

In the set of relevant papers, we identified two sub-groups: methodology papers mentioning cohort studies and empirical studies reporting cohort studies. We will discuss them next.

*4.1.1 Methodology Papers that Mention Cohort Studies.* Several sources mention field studies which seem to have commonalities with cohort studies. Zelkowitz and Wallace [32] describe field studies as a subtype of observational studies which "examine data collected from several projects simultaneously" that can be used for making comparisons. This high-level definition could be compatible with that of a cohort study. More recently, Singer et al. [26] define field studies broadly as just studies done with real practitioners as they work in their normal context. Nevertheless, some of the described data collection techniques could be used for conducting cohort studies in SE.

Easterbrook et al. [5] briefly mention cohort studies when describing survey research. In their definition of cohort studies, they are used to "track changes over time for a group of participants". Therefore, the definition is closer to an observational case-study than to an epidemiology cohort study. Even Eastbrook et al.'s definition contains the temporal aspect of a cohort study, it lacks the systematical comparison of groups with different treatments.

The definitions of observational study and cohort study vary significantly between SE and epidemiology. Specifically, the term "cohort study" has not been defined properly in SE.

*4.1.2 Empirical Studies Running Cohort Studies.* We identified only one paper describing itself as a cohort study. Fucci et al. [6] investigate if applying Test-Driven Development affects the quality of software or developers productivity. However, despite using the term cohort study in the paper title, it is not a cohort study in the sense epidemiologists use it. Unlike cohort studies, in this study cohorts are randomized, treatments are manipulated, and there is no control group.

While there were no other papers using the term cohort study, we identified papers that, not filling the definition of a cohort study according to epidemiology, share similar characteristics with them. White and Coffrey [29] introduce an executive SE program in which the students complete their Master's degree in one calendar year. The students in the program form a cohort that do their studies together. Unlike cohort studies, this study contains a very limited comparison between the students inside and outside the cohort.

Godfrey et al. [8] follow a group of students and investigate their reasons for retention and attrition. They perform "cohort analysis" in order to determine differences between the different groups where they consider factors like gender and being a domestic vs international student. Unlike cohort studies, specific exposures are not investigated, as the main goal is not only to define whether there are significant differences but also to map why some students did not graduate.

In view of the results of our literature search, we can conclude that **there are no current works conducting cohort studies** (although some claim they are performing cohort studies), **nor are they properly allocated from the methodological perspective**.

## 4.2 Guidelines for Empirical Studies

The SE community has proposed guidelines for conducting the different kinds of studies that are commonly used.

Jedlitschka et al. [13] have described a guideline for reporting **controlled (and quasi-)experiments**. It aims at helping researchers to write clear and well structured papers by proposing a structure for them. There are also guidelines for conducting experiments. Juristo and Moreno [17] and Wohlin et al. [31] illustrate how to conduct experiments in SE covering all aspects from the basic concepts to data analysis.

Runeson and Höst [25] have written a guideline for conducting and reporting **case studies** in SE. It describes the complete process for planning and conducting a case-study as well as for analyzing the data. It also provides a structure for reporting the study and checklists for conducting and reading case studies.

Kitchenham et al. [19] have defined a general guideline for conducting **SLRs** in SE. Their guideline covers all aspects of SLR from defining the term to planning, conducting, and using a structured way of reporting them. In addition, Wohlin [30] has written a guideline for replicating SLRs and applying the snowballing technique for finding papers. Garousi et al. [7] have created a guideline for conducting MLRs in SE[2]. The guideline covers the whole process for performing a MLR.

Carver [2] has proposed a preliminary guideline for reporting **replications of experiments**. In addition, Juristo and Gómez [16] investigate the existing replications types in other disciplines and provide guidelines to identify replication types in SE. It is not a guideline itself, but it complements Carver's guideline.

In view of the results, we can conclude that **there are no guidelines for conducting cohort studies in SE**.

## 4.3 Guidelines for MSR Studies

Nagappan et al. [22] have proposed a guideline to address the diversity of the projects from which to get the data set to be used in MSR studies. The authors discuss the principles for choosing a set of projects with a higher coverage: overall population of projects (universe), the relevant aspects of the projects (space), and a list of similarity rules for the projects (configuration), and propose a technique for assessing the coverage of selected projects.

[2]Multivocal literature reviews (MLRs) are a type of SLR that enable the inclusion of the gray literature. This provides a broader view on the selected research questions.

In view of the results, we can conclude that **there is no guideline for conducting cohort studies in the MSR field**.

# 5 SETTING THE FOUNDATIONS FOR COHORT STUDIES IN MSR (AND SE)

Controlled experiments are the only empirical method capable of obtaining cause-effect relationships. In this section we review the principles of experimentation that allow the identification of causality, and identify which ones violate cohort studies (Section 5.1). We next discuss how this affects the internal validity of a cohort study (Section 5.2). Finally, we identify the mechanism and associated techniques that cohort studies use to tackle this, and discuss how a process for conducting cohort studies based on the existing SE process for experiments should incorporate them.

## 5.1 Governing Principles

Four principles govern controlled experiments [20]. *Treatment design*—also known as manipulation of the independent variable—consists of choosing (designing) the proper treatments according to the formulated research hypothesis. *Local control* is a series of actions through which the researcher controls experimental protocol, selection of experimental units (so that they are uniform), blocking (to ensure parity on all treatments), choice of experimental design, and measurement of covariates. *Randomization* allows researchers to proceed as if the observations are independent and constitute a random sample from a normally distributed population. *Replication* implies that each treatment is applied independently to each of two or more experimental units.

Table 1 shows whether these principles are tackled by the most common primary study types performed in SE (controlled experiments, quasi-experiments and case studies), and to what extent they can lead to the identification of causality. Cohort studies have also been included. It is important to note that the borderline between different types of study is not always clear cut [31].

**Table 1: Principles tackled by different study types**

| Principle | Controlled experiment | Quasi-experiment | Cohort study | Case study |
|---|---|---|---|---|
| Treatment design | Yes | Yes | No | No |
| Local control | Yes | Yes | Yes | No |
| Randomization | Yes | No | No | No |
| Replication | Yes | Yes | Yes | No |
| Causality | Easy | Moderate | Difficult | Impossible |

According to Table 1, quasi-experiments lack randomization, which makes the difficulty of obtaining causality moderate, while case studies lack all of them, thus preventing them from obtaining causality. The two principles that cohort studies lack are [24]: treatment design and randomization. Additionally, local control is exercised differently from experiments. This makes it difficult to elucidate cause/effect relationships. In a controlled experiment:

- The assignment of treatments to experimental units is controlled by the experimenter—by means of randomization, who ensures that units receiving different treatments are comparable. In cohort studies, the researcher cannot control

this assignment and therefore a different technique must be used to ensure that the units that received the different treatments are similar (this is discussed in Section 5.3).
- The experimenter manipulates (designs) the treatments. In a cohort study, the researcher can only measure variables as they naturally occur.
- The environment is tightly controlled by design. In cohort studies the environment is carefully chosen.

The existence of causality requires that three conditions are met [24]: empirical association (covariation) between the independent and dependent variables, temporal precedence of the independent variable, and nonspuriousness (control for third variables). In a cohort study the third condition for causality is not guaranteed and must be achieved.

It is worth noting that epidemiologists have used the methodology for a long time and therefore based on practical experience they know how to consider the correct aspects while making a study in their field. Therefore, it is expected that adapting the methodology to ESE in a way that is generalizable and unbiased will take time.

## 5.2 The Confounding (Internal) Validity Threat

For the reasons mentioned above, internal validity is weaker in cohort studies than in experiments. More precisely, cohort studies might suffer from *confounding bias* [24]. It is a distortion that modifies the association between the independent and the dependent variable because there exists another (confounding) variable. Figure 4 represents the principle of confounding. For a variable to be a confounder it must [10]: (1) have an association with the independent variable, (2) be associated with the dependent variable, and (3) must not be an effect of the independent variable.

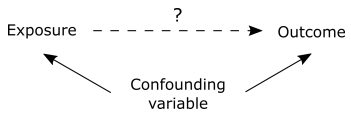


**Figure 4: The principle of confounding.**

Failing to control for confounding variables can cause an over- or under-estimate of the observed association between the independent and the dependent variable. The confounder makes the exposure more likely and in some way independently modifies the outcome, making it appear that there is an association between the exposure and the outcome when there is none, or masking a true association. Therefore, the mechanism that cohort studies use to obtain causality is avoiding confounding bias.

## 5.3 Dealing with the Confounding Threat

The techniques that cohort studies use to articulate the mechanism of avoiding the confounding threat are shown in Table 2. Figure 5 visualizes our proposal of how the process for conducting controlled experiments in SE should be modified to incorporate the techniques in Table 2, and obtain a process for cohort studies. Green steps are done in both study types—but differently, while red and blue ones are done only in experiments and cohort studies respectively.

**Table 2: Techniques to deal with the confounding threat**

| Phase | Controlled experiment | Cohort study |
|---|---|---|
| Scoping | | Elaborate theories |
| Planning | Manipulation<br>Randomization | Restriction<br>Matching<br>Stratification |
| | | Confounding validity |
| | Control environment | Measure environment |
| Operation | Execution | Observation |
| Analysis & Interpretation | | Statistical adjustment<br>Sensitivity analysis<br>Criticism |

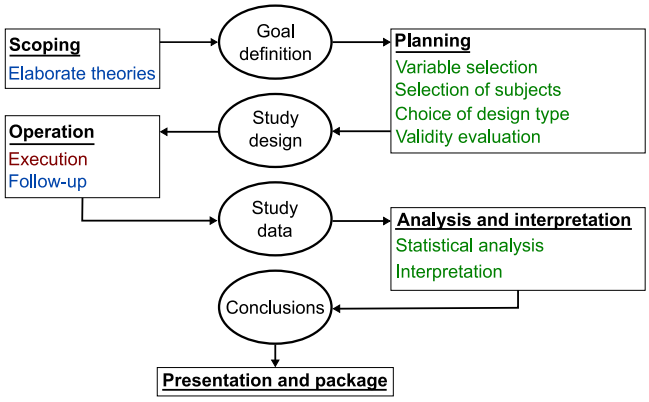


**Figure 5: Foundations of a process for cohort studies.**

During the scoping phase, a key issue in cohort studies is establishing theories. As noted by Cochran [3], what can be done in observational studies to clarify the step from association to causation is "making your theories elaborate". This means that when constructing a causal hypothesis one should envisage as many different consequences of its truth as possible, and plan the studies to discover whether each of these consequences is found to hold. This multi-phasic attack is one of the most potent weapons in cohort studies.

During the planning phase, in cohort studies, the researchers cannot manipulate the independent variable (design the treatments). Instead they can only select the ones applied by the chosen subjects (restricting the non interesting ones). At the same time, subjects cannot be assigned to treatments randomly. Since the validity of a cohort study lies in the comparability of the treatment groups, it is necessary that the subjects are as similar as possible. This way, potentially confounding variables are identified and techniques such as matching and stratification are used to counteract their effect. Finally, the confounding validity threat needs to be thoroughly assessed.

The difference in operation between an experiment and a cohort study is that while an experiment is run (or executed), a cohort study is observed (follow-up).

During the analysis phase, the possible bias due to confounders needs to be addressed. For this purpose, two new types of analyses have to be performed after the regular one. The first one addresses

overt bias, and consists of using techniques that adjust for the confounding variables identified during planning (e.g. ANCOVA). The second one addresses hidden bias due to non-identified confounders, and is named sensitivity analysis.

Finally, during the interpretation stage, criticism must be exercised. This is of great importance. What is scientifically plausible must be distinguished from what is just logically possible [24]. In words of Cochran [3], the first critic of a cohort study should be its author. When summarizing the results of a study that shows an association consistent with the causal hypothesis, the researcher should always list and discuss all alternative explanations of the results (including different hypotheses and biases in the results) that occur to her.

## 6 THE ROADMAP

In order for MSR (and SE) to adopt cohort studies as a methodology, it is necessary to further develop the foundations established in Section 5. Our proposed roadmap includes the following steps:

(1) Understand how cohort studies are run in epidemiology.
(2) Delineate what kind of SE research questions cohort studies may be best suited for.
(3) Define cohort studies in MSR by adapting the concepts, methods and techniques used in epidemiology to the characteristics of MSR.
(4) Establish a process to conduct cohort studies in MSR starting from the one defined for controlled experiments.
(5) Improve the proposed process by applying it to previous MSR studies and comparing the obtained results.
(6) Elaborate guidelines for cohort studies in MSR.
(7) Explore other SE contexts, apart from MSR, where cohort studies can be performed.

## 7 CONCLUSIONS

The relationships between variables identified in MSR studies are not causal because controlled experiments cannot be run. In these situations, epidemiologists conduct cohort studies.

We believe that cohort studies are a promising approach that could help MSR researchers to improve the level of evidence of their findings, eventually arriving at the identification of possible causal relations.

In this work, we envision the use of cohort studies in the MSR field—and eventually in ESE, identify the main mechanism they use and the techniques that articulate it., and outline a roadmap towards its adoption. This rodadmap includes understanding how cohort studies are run in epidemiology, adapting them to the MSR context, elaborate guidelines to conduct them, and explore other SE situations where they could be used.

## REFERENCES


[1] Gabriele Bavota, Andrea De Lucia, Massimiliano Di Penta, Rocco Oliveto, and Fabio Palomba. 2015. An experimental investigation on the innate relationship between quality and refactoring. *Journal of Systems and Software* 107 (2015), 1 – 14.

[2] Jeffrey C Carver. 2010. Towards reporting guidelines for experimental replications: A proposal. In *1st international workshop on replication in empirical software engineering*, Vol. 1. Citeseer, 1–4.

[3] William G. Cochran. 1965. The planning of observational studies of human populations (with discussion). *Journal of the Royal Statistical Society* Series A, 128 (1965), 134–155.

[4] Maria Elisa Di Cicco, Vincenzo Ragazzo, and Tiago Jacinto. 2016. Mortality in relation to smoking: the British Doctors Study. *Breathe* 12 (3 2016), 275–276.

[5] Steve Easterbrook, Janice Singer, Margaret-Anne Storey, and Daniela Damian. 2008. Selecting empirical methods for software engineering research. In *Guide to advanced empirical software engineering*. Springer, 285–311.

[6] Davide Fucci, Simone Romano, Maria Teresa Baldassarre, Danilo Caivano, Giuseppe Scanniello, Burak Turhan, and Natalia Juristo. 2018. A longitudinal cohort study on the retainment of test-driven development. In *12th International Symposium on Empirical Software Engineering and Measurement*. 1–10.

[7] Vahid Garousi, Michael Felderer, and Mika V. Mäntylä. 2019. Guidelines for including grey literature and conducting multivocal literature reviews in software engineering. *Information and Software Technology* 106 (2019), 101 – 121.

[8] Elizabeth Godfrey, Tim Aubrey, and Robin King. 2010. Who leaves and who stays? Retention and attrition in engineering education. *engineering education* 5, 2 (2010), 26–40.

[9] Sherita Golden and Eric Bass. 2013. Validity of Meta-analysis in Diabetes: Meta-analysis Is an Indispensable Tool in Evidence Synthesis. *Diabetes care* 36 (10 2013), 3368–3373.

[10] Leon. Gordis. [n.d.]. *Epidemiology* (fifth edition. ed.). Elsevier Saunders, Philadelphia.

[11] David A. Grimes and Kenneth F. Schulz. 2002. An overview of clinical research: The lay of the land. *Lancet* 359, 9300 (2002), 57–61.

[12] David A. Grimes and Kenneth F. Schulz. 2002. Cohort studies: Marching towards outcomes. *Lancet* 359, 9303 (2002), 341–345.

[13] Andreas Jedlitschka, Marcus Ciolkowski, and Dietmar Pfahl. 2008. Reporting Experiments in Software Engineering. In *In Guide to Advanced Empirical Software Engineering*. Springer, 201–228.

[14] Ross D. Jeffery and Lawrence G. Votta. 1999. Empirical software engineering. *IEEE Transactions on Software Engineering* 25, 4 (1999), 435–437.

[15] Natalia Juristo. 2018. Use and Misuse of the term experiment in the software repositories research. (2018). 12th International Conference on Predictive Models and Data Analytics in Software Engineering (PROMISE).

[16] Natalia Juristo and Omar S Gómez. 2010. Replication of software engineering experiments. In *Empirical software engineering and verification*. Springer, 60–88.

[17] Natalia Juristo and Ana M Moreno. 2013. *Basics of software engineering experimentation*. Springer Science & Business Media.

[18] Foutse Khomh, Massimiliano Di Penta, Yann-Gaël Guéhéneuc, and Giuliano Antoniol. 2012. An Exploratory Study of the Impact of Antipatterns on Class Change- and Fault-Proneness. *Empirical Softw. Engg.* 17, 3 (2012), 243–275.

[19] Barbara Kitchenham and Stuart Charters. 2007. Guidelines for performing Systematic Literature Reviews in Software Engineering.

[20] Robert O. Kuehl. 2000. *Design of experiments: statistical principles of research design and analysis* (2 ed.). BRROKS/COLE CENGAGE Learning.

[21] Tim Menzies. 2016. Correlation is not causation (or, when not to scream "Eureka!"). In *Perspectives on Data Science for Software Engineering*, Tim Menzies, Laurie Williams, and Thomas Zimmermann (Eds.). Morgan Kaufmann, 327–330.

[22] Meiyappan Nagappan, Thomas Zimmermann, and Christian Bird. 2013. Diversity in software engineering research. In *9th joint meeting on foundations of software engineering*. 466–476.

[23] Fabio Palomba, Gabriele Bavota, Massimiliano Di Penta, Fausto Fasano, Rocco Oliveto, and Andrea De Lucia. 2018. On the Diffuseness and the Impact on Maintainability of Code Smells: A Large Scale Empirical Investigation. *Empirical Softw. Engg.* 23, 3 (June 2018), 1188–1221.

[24] Paul R. Rosenbaum. 2002. *Observational studies* (2 ed.). Springer.

[25] Per Runeson and Martin Höst. 2009. Guidelines for conducting and reporting case study research in software engineering. *Empirical Software Engineering* 14 (2009), 131–164. Issue 2.

[26] Janice Singer, Susan E Sim, and Timothy C Lethbridge. 2008. Software engineering data collection for field studies. In *Guide to Advanced Empirical Software Engineering*. Springer, 9–34.

[27] Jae W. Song and Kevin C. Chung. 2010. Observational Studies: Cohort and Case-Control Studies. *Plastic and Reconstructive Surgery* 126, 6 (2010), 2234–2242.

[28] Michele Tufano, Fabio Palomba, Gabriele Bavota, Rocco Oliveto, Massimiliano Di Penta, Andrea De Lucia, and Dennis Poshyvanyk. 2017. When and Why Your Code Starts to Smell Bad (and Whether the Smells Go Away). *IEEE Transactions on Software Engineering* 43, 11 (2017), 1063–1088.

[29] Laura J White and John Coffey. 2011. The design and implementation of an innovative online program for a master of science degree in Computer Science—Software Engineering specialization. In *24th Conference on Software Engineering Education and Training (CSEE&T)*. 257–265.

[30] Claes Wohlin. 2014. Guidelines for snowballing in systematic literature studies and a replication in software engineering. In *18th international conference on evaluation and assessment in software engineering*. 1–10.

[31] Claes Wohlin, Per Runeson, Martin Höst, Magnus C Ohlsson, Björn Regnell, and Anders Wesslén. 2012. *Experimentation in software engineering*. Springer Science & Business Media.

[32] Marvin V. Zelkowitz and Dolores R. Wallace. 1998. Experimental models for validating technology. *Computer* 31, 5 (1998), 23–31.