ohort Studies for Mining Software Repositories

Nyyti Saarimäki[1], Sira Vegas[2], Valentina Lenarduzzi[3], Davide Taibi[3], Mikel Robredo[3]
[1] University of Luxembourg — [2] Universidad Politécnica de Madrid — [3] University of Oulu
nyyti.saarimaki@uni.lu; svegas@fi.upm.es; valentina.lenarduzzi@oulu.fi;
davide.taibi@oulu.fi; mikel.robredomanero@oulu.fi

ABSTRACT

Mining Software Repositories studies have become increasingly popular over the years. However, a notable limitation is that they report correlational relationships rather than establishing causation. In contrast, certain disciplines (e.g. epidemiology) have developed specific methods to address this limitation. The goal of this tutorial is to introduce participants to one such method: cohort studies. By the end of the tutorial, participants will be familiar with the steps and techniques involved in designing and analyzing cohort studies.

ACM Reference Format:
Nyyti Saarimäki[1], Sira Vegas[2], Valentina Lenarduzzi[3], Davide Taibi[3], Mikel Robredo[3]. 2024. Cohort Studies for Mining Software Repositories. In *21st International Conference on Mining Software Repositories MSR '24), April 15–16, 2024, Lisbon, Portugal.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3643991.3649103

1 AIMS AND OBJECTIVES

This tutorial seeks to enhance the exploration of cause-and-effect relationships in the field of Mining Software Repositories (MSR) by providing guidance on conducting cohort studies. This goal will be accomplished through three main objectives: 1) understanding the difference between correlation and causation in MSR studies, and its implications; 2) learning how to use cohort studies in MSR; and 3) applying the cohort study methodology within the MSR field.

2 AUDIENCE AND BACKGROUND

This tutorial is primarily designed for SE researchers, ranging from PhD students to senior researchers, who are interested in exploring cause-and-effect relationships in the context of MSR studies. A basic understanding of SE and MSR is required.

3 RELEVANCE

Controlled experiments are considered the gold standard for studying cause-and-effect relationships [6]. However, in the context of MSR, they are impossible to run, as these studies rely on existing data from software repositories and monitoring tools to understand the relationships between variables [4]. Consequently, many conducted MSR studies are correlational, with researchers acknowledging that the observed relationships may not imply causation [1, 5, 8].

Relying solely on correlational research does not benefit the MSR community. Instead, it would be advantageous to employ research methodologies that allow for obtaining a higher level of evidence and, eventually, to gain an understanding of causality. This approach would contribute to the maturation of the field.

Other scientific fields have developed research methodologies to study cause-and-effect relationships using only existing data. This is the case for epidemiology[1]. Often, epidemiologists cannot conduct controlled experiments and must rely on observational data. A notable achievement in this field was establishing the cause-effect link between tobacco and lung cancer [3].

One of the methods employed by epidemiologists to investigate cause-and-effect relationships with observational data is the use of cohort studies. These studies examine whether a specific factor (or factors) causes an outcome over time by comparing the outcomes of two or more groups of study subjects with varying levels of the factor. Applying cohort studies in MSR research could assist the community in exploring causality using the same data currently utilized to conduct correlational studies.

Since the inception of MSR studies, their popularity has steadily increased. Nowadays, they constitute a significant percentage of the empirical studies conducted in SE, having being adopted in industry [2, 9]. Thus, the tutorial would be of interest to a diverse audience, including individuals from both academia and industry.

4 OUTLINE

- **Introduction** Examines the distinctions between correlation and causation, and discusses the principles and mechanisms necessary for establishing a causal relationship.
- **Observational studies** Introduces various types of studies utilized in epidemiology, with a specific focus on the steps involved in conducting a cohort study in MSR.
- **Planning a cohort study:** Presents the steps for planning a cohort study (hypothesis formulation, context and subject selection, variables identification, and instrumentation) accompanied by an illustrative example.
- **Avoiding extraneous variables:** Describes the techniques for designing a cohort study (restriction and matching) supported by an example.
- **Exploring extraneous variables:** Covers the strategy for analyzing and interpreting a cohort study, featuring a practical example.

[*] Part of this work was done while the author was working at Tampere University

MSR '24, April 15–16, 2024, Lisbon, Portugal
© 2024 Association for Computing Machinery.
ACM ISBN 979-8-4007-0587-8/24/04… 15.00
https://doi.org/10.1145/3643991.3649103

[1] A branch of medicine that studies diseases and other health-related factors in specified populations (neighborhood, school, country, global, etc.) [7].

ACKNOWLEDGMENTS

Grant PID2022-137846NB-I00 funded by MCIN/AEI/10.13039/501100011033, by "ERDF A way of making Europe".

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).

[2] Tingting Bi, Xin Xia, David Lo, John Grundy, and Thomas Zimmermann. 2020. An empirical study of release note production and usage in practice. *IEEE Transactions on Software Engineering* (2020).

[3] 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.

[4] Ahmed E Hassan. 2008. The road ahead for mining software repositories. In *2008 Frontiers of Software Maintenance*. IEEE, 48–57.

[5] 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.

[6] Marcia L Meldrum. 2000. A brief history of the randomized controlled trial: From oranges and lemons to the gold standard. *Hematology/oncology clinics of North America* 14, 4 (2000), 745–760.

[7] Miquel Porta. 2014. *A dictionary of epidemiology.* Oxford university press.

[8] 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.

[9] Jie Zhang, Feng Li, Dan Hao, Meng Wang, Hao Tang, Lu Zhang, and Mark Harman. 2019. A Study of Bug Resolution Characteristics in Popular Programming Languages. *IEEE Transactions on Software Engineering* (2019).