



## Introduction to Causal Inference for Software Engineering

Downloaded from: <https://research.chalmers.se>, 2026-10-10 16:30 UTC

Citation for the original published paper (version of record):

Siebert, J., Frattini, J., Heyn, H. et al (2026). Introduction to Causal Inference for Software Engineering. Fse Companion 2026 Proceedings of the 34th ACM International Conference on the Foundations of Software Engineering: 52-53. <http://dx.doi.org/10.1145/3803437.3804898>

N.B. When citing this work, cite the original published paper.

# Introduction to Causal Inference for Software Engineering

Julien Siebert

julien.siebert@iese.fraunhofer.de  
Fraunhofer IESE  
Kaiserslautern, Germany

Hans-Martin Heyn

heyne@chalmers.se  
Chalmers University of Technology and Gothenburg  
University  
Sweden

Julian Frattini

julian.frattini@chalmers.se  
Chalmers University of Technology and Gothenburg  
University  
Sweden

Roberto Pietrantuono

roberto.pietrantuono@unina.it  
University of Naples Federico II  
Naples, Italy

## Abstract

The ever-increasing use of data-driven models (whether classical machine learning or novel deep learning models, such as LLMs) to automate software engineering (SE) is accentuating the need to distinguish between causal effects and spurious correlations. While the application of causal methods in SE is gradually attracting more attention, it is not yet widely used. This 90 min. tutorial provides an introduction to causal inference techniques for observational data based on causal graphs and its application to SE. The goal is to enable SE researchers to apply such methods to their own use cases and streamline the knowledge and use of causal inference within the SE community. Causal inference methods focus on the question of identifying and estimating the causal impact of a variable of interest on another one even when the data was not collected in a randomized controlled experiment (i.e., using observational data). Besides the theory, this tutorial will also showcase how such methods have been applied in SE and it will give pointers to libraries in Python and R.

## CCS Concepts

• **Software and its engineering**; • **Computing methodologies** → *Artificial intelligence*; • **General and reference** → *Cross-computing tools and techniques*;

## Keywords

Causality, Software Engineering, Causal Inference

### ACM Reference Format:

Julien Siebert, Julian Frattini, Hans-Martin Heyn, and Roberto Pietrantuono. 2026. Introduction to Causal Inference for Software Engineering. In *34th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (FSE Companion '26)*, July 05–09, 2026, Montreal, QC, Canada. ACM, New York, NY, USA, 2 pages. <https://doi.org/10.1145/3803437.3804898>



This work is licensed under a Creative Commons Attribution 4.0 International License. FSE Companion '26, Montreal, QC, Canada  
© 2026 Copyright held by the owner/author(s).  
ACM ISBN /2026/07  
<https://doi.org/10.1145/3803437.3804898>

## 1 Causal Inference

Causal inference focuses on estimating causal effects of an action (a treatment  $T$ ) on a given observed system (an outcome  $Y$ ) from observational data (i.e., data that was not generated during randomized controlled trials) [7, 11, 17]. The fundamental problem with observational data is that we cannot have the same subject being treated and not being treated at the same time (following Pearl's notation and assuming a binary treatment, we cannot observe  $P(Y|\text{do}(T = 0))$  and  $P(Y|\text{do}(T = 1))$  at the same time) and the effect has to be estimated from available data (i.e., from  $P(Y|T = 0)$  and  $P(Y|T = 1)$ ). The challenge with observational data, assuming that all relevant variables have been captured, is that it can contain correlations induced by confounding factors that can bias the estimation of causal effects. For this reason, researchers in the field of causal inference have developed methods to reduce the impact of confounding factors and separate spurious correlations from causal effects [3, 9]. One of such approach, is the use of graphical causal models together with the application of the do-calculus by Pearl et al. [11, 12].

Graphical causal models are a way to represent hypotheses about causality in a given system. They take the form of a graph<sup>1</sup> consisting of nodes representing the relevant variables (potentially measurable) and links representing a direct causal effect between two variables. For example, if a treatment  $T$  is supposed to have a causal effect on an outcome  $Y$ , then both  $T$  and  $Y$  are represented as nodes and a link goes from  $T$  and points towards  $Y$ . Graphical models are interesting for different reasons. First, they make hypotheses about variables under study and their causal effects explicit. Second, they are used for identifying structures that lead to spurious correlations and need to be handled before estimation.

The methodological process behind Pearl's approach consists of four major steps: modeling, identification, estimation and refutation [12, 13]. The first step (modeling) aims to make causal assumptions explicit through the a graphical causal model. The second step (identification) aims to find out whether a given causal effect can be estimated using the available data. Here the graphical structure is used to analyze for the presence of sources of spurious correlations. This requires identifying structures such as confounders, mediators,

<sup>1</sup>Causal graphs are by definition directed acyclic graph and are sometimes referred to (causal) DAG.

or colliders to understand which variable(s) should be controlled for [2]. The third step (estimation) aims to effectively estimate the identified causal effect based on the available data and the previously gained knowledge of which variables should be adjusted for. Estimation can be done using a variety of methods, from simple linear regression to more complex approaches such as T-learner, X-learner, double machine learning or causal random forests<sup>2</sup>. The last step (refutation) consists in trying to falsify the causal estimate<sup>3</sup>.

## 2 Tutorial aims and objectives

This tutorial provides an introduction to causal inference techniques for observational data based on causal graphs. The main goal is to provide researchers with a solid understanding of the basis of these methods, relevant reading material and libraries to use as well as a short overview of their application in SE (see [6, 14]). No special background knowledge in Statistics is required. Intended audience are researchers interested in drawing causal conclusions from quantitative data collected with observational research methods. Table 1 gives an outline of the topics covered by the tutorial, with approximate timing.

|                                                                                                                                                                                                      |        |
|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------|
| Introduction (material and exercises access). Motivation from a SE perspective.                                                                                                                      | 10 min |
| Modeling causal effects with causal graphs. Examples from SE research papers (most probably from domains such as Requirements Engineering, Testing and Fault Localization [5, 8, 10]).               | 15 min |
| Basic causal graphical structure and d-separation.                                                                                                                                                   | 15 min |
| Break.                                                                                                                                                                                               | 10 min |
| Identification, adjustment sets and conditional independencies. Exercises with Dagitty <sup>4</sup> and application to previous SE examples.                                                         | 15 min |
| Causal effect estimation and short overview in causal machine learning.                                                                                                                              | 15 min |
| Conclusion: methods not explicitly relying upon graphs, counterfactual and mediation analysis, next steps in your causal journey with pointers to books, online Lectures and other reading material. | 10 min |

**Table 1: Outline of the tutorial**

The participants will learn how to specify and draw causal graphs (both manually and using automated tools such as Dagitty<sup>5</sup>), what these graphs can tell us about how the data should behave if the causal assumptions underlying the graph hold true (d-separation), and how they can be used to separate spurious correlations from causal effects (do-calculus rules of identification). The participants will get a glimpse of the different estimation techniques in the field of causal machine learning and learn how to apply them in Python and R. The theory will be illustrated using examples of how causal methods have been applied in the field of software engineering (SE) [5, 8, 10].

<sup>2</sup>The interested reader can refer to Chapter 7 in [1] or Part 2 of [9] for an overview of estimation methods.

<sup>3</sup>See refutation methods implemented in the DoWhy library [https://www.pywhy.org/dowhy/v0.8/user\\_guide/effect\\_inference/refute.html](https://www.pywhy.org/dowhy/v0.8/user_guide/effect_inference/refute.html) and [13].

<sup>5</sup><https://dagitty.net/> [16]

The intended duration is 90 minutes and the format is a Hands-On Session with interactive exercises. This tutorial is based on previous lectures at the university of Napoli from Roberto Pietrantuono, a guest lecture at the university of TU Darmstadt from Julien Siebert, and the tutorial "Applied Statistical Causal Inference in Requirements Engineering" given at the RE'25 conference in Valencia by Julian Frattini<sup>6</sup> [4]. This material is adapted to fit a broader audience of SE researchers.

Causal analysis methods, whether based on graphs or not, are relevant for empirical work and for distinguishing between causal effects and spurious correlations. In fact, the 2021 Nobel Prize in Economics was awarded for "answering causal questions using observational data" (see <https://www.nobelprize.org/uploads/2021/10/advanced-economicsciencesprize2021.pdf>). These methods are also useful for discussing the impact of policies on changes in SE practices. Methods based on causal graphs are relatively new and not widespread in the SE community (see [6, 14, 15] and the CauSE2025 Workshop in Proceedings of the 33rd ACM International Conference on the Foundations of Software Engineering (FSE2025)) and this tutorial also aims to strengthen and streamline causal inference in SE research and applications.

## References

- [1] N. Brady. *Introduction to Causal Inference from a Machine Learning Perspective (Course Lecture Notes)*.
- [2] C. Cinelli, A. Forney, and J. Pearl. A crash course in good and bad controls. *Sociological Methods & Research*, page 00491241221099552, 2022.
- [3] S. Cunningham. *Causal Inference: The Mixtape*. Yale University Press, 2021.
- [4] J. Frattini, H.-M. Heyn, R. Feldt, and R. Torkar. Tutorial debriefing: Applied statistical causal inference in Requirements Engineering, 2025.
- [5] L. Giamattei, A. Guerriero, R. Pietrantuono, and S. Russo. Causality-driven testing of autonomous driving systems. *ACM Transactions on Software Engineering and Methodology*, 33(3):1–35, 2024.
- [6] L. Giamattei, A. Guerriero, R. Pietrantuono, and S. Russo. Causal reasoning in software quality assurance: A systematic review. *Information and Software Technology*, 178:107599, 2025.
- [7] J. Hernán, M.A.; Robins. *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC, 2020.
- [8] H.-M. Heyn, Y. Mao, R. Weiß, and E. Knauss. Causal models for specifying requirements in industrial ML-based software: A case study. *Journal of Systems and Software*, page 112691, 2025.
- [9] N. Huntington-Klein. *The Effect: An Introduction to Research Design and Causality*. CRC Press, 2021.
- [10] Y. Küçük, T. A. Henderson, and A. Podgurski. Improving fault localization by integrating value and predicate based causal inference techniques. In *2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE)*, pages 649–660. IEEE, 2021.
- [11] J. Pearl, M. Glymour, and N. P. Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016.
- [12] J. Pearl and D. Mackenzie. *The Book of Why*. Penguin Books, 2019.
- [13] A. Sharma and E. Kiciman. *Dowhy: An end-to-end library for causal inference*. *arXiv preprint arXiv:2011.04216*, 2020.
- [14] J. Siebert. Applications of statistical causal inference in Software Engineering. *Information and Software Technology*, 159:107198, 2023.
- [15] J. Siebert, A. Trendowicz, G. Gössler, H. Washizaki, M. Kläs, and M. Shepperd. Special issue on causal modeling and inference in SE, 2025.
- [16] J. Textor, B. Van der Zander, M. S. Gilthorpe, M. Liśkiewicz, and G. T. Ellison. Robust causal inference using directed acyclic graphs: the R package 'dagitty'. *International journal of epidemiology*, 45(6):1887–1894, 2016.
- [17] L. Yao, Z. Chu, S. Li, Y. Li, J. Gao, and A. Zhang. A survey on causal inference. *ACM Trans. Knowl. Discov. Data*, 15(5), may 2021.

<sup>6</sup>Tutorial material available at <https://doi.org/10.5281/zenodo.17649808>