Biometrical Journal

WILEY-VCH

Biometrical Journal

DISCUSSANT'S COMMENTS OPEN ACCESS

Hazards, Causality, and Practical Relevance of Collider Effects – Comment on Beyersmann et al. "Hazards Constitute Key Quantities for Analyzing, Interpreting and Understanding Time-to-Event Data"

Ralf Bender[1,2] | Lars Beckmann[1]

[1]Department of Medical Biometry, Institute for Quality and Efficiency in Health Care (IQWiG), Cologne, Germany | [2]Faculty of Medicine, University of Cologne, Cologne, Germany

Correspondence: Ralf Bender (Ralf.Bender@iqwig.de)

Received: 13 May 2025 | **Revised:** 31 July 2025 | **Accepted:** 2 August 2025

Keywords: causality | collider effect | estimand | hazard ratio | hazards

ABSTRACT

Hazards constitute key quantities for analyzing, interpreting, and understanding time-to-event data. Hazards and corresponding effect measures, such as the hazard ratio from the Cox proportional hazards model, have a valid causal interpretation if the hazard function is considered as a function in time rather than hazards at specific time points. In this comment, we would like to add two points: (1) The hazard ratio is also a useful population-level estimand with a valid causal interpretation. (2) Empirical evidence shows that problematic situations, which could occur in theory due to strong heterogeneity, are usually avoided in typical randomized controlled trials.

The article by Beyersmann et al. (2025) takes up the current discussion of whether hazards and the hazard ratio (HR) estimated by the Cox proportional hazards model have a valid causal interpretation. The features of survival data, censoring, hazards, hazard functions, and HRs are explained and discussed. Arguments are presented that a causal interpretation of hazards is possible if the hazard function is considered as a function in time rather than hazards at specific time points.

Examples from the framework of benefit assessment are discussed, leading to interesting insights into the issue of problematic analyses of adverse events by means of naive contingency tables in dossiers assessed by the Institute for Quality and Efficiency in Health Care (Bender et al. 2016). New empirical evidence is presented that similar misleading analyses and misinterpretations of results are existing not only in submitted dossiers in Germany but also in the international medical literature. Additionally, theoretical explanations are provided showing the importance of hazard functions for the estimation of causal effects in the case of time-to-event data.

We congratulate Jan, Claudia, and Martin on this useful paper, because it is very important for biostatistical practice to underline that hazards and HRs are key quantities for time-to-event data with censoring and to clarify in which way they have a valid causal interpretation in randomized controlled trials (RCTs).

We like to add two related discussion points that Beyersmann et al. (2025) mostly have avoided: (1) the issue of population-level versus individual-level estimands and (2) empirical evidence regarding the effect of strong heterogeneity, which could lead to a collider effect in theory.

The approval and benefit assessment of new drugs requires conclusions on their performance at the population level. Beyersmann et al. (2025) already noted at the end of the paper that population-level hazards do not equal individual-level hazards in general and that they applied the population-level view to avoid further subtleties in the debate. We think that it is important to underline that the HR is a useful population-level estimand, which has a valid causal interpretation in RCTs as shown by Fay and Li (2024). In causal inference, many statistical approaches are based on individual contrasts of potential outcomes, and correspondingly, averages of individual contrasts are used for effect estimation. We like to add that such individual-level estimands do not represent the natural estimand in RCTs with parallel group design, because individual contrasts are not directly observable. In trials with comparison of two groups, the natural estimand is the population-level estimand, that is, the contrast of the averages of the two groups. Difference-in-expectation estimands, such as mean and risk differences, are both population-level and individual-level estimands (Fay and Li 2024), but this is not the case for the HR. Nevertheless, as shown by Fay and Li (2024), if Cox models with more and more baseline covariates are used, the population-level estimand approaches more and more to an individual-level estimand. As already noted by Beyersmann et al. (2025), the issue of population-level versus individual-level estimands is subtle. In the case of approximately constant treatment effects over time, there is no difference in practical applications whether an estimated treatment effect is interpreted in a population-level or individual-level view, because this has no consequence in decision-making processes such as drug approval or benefit assessment. In any case, the focus of drug approval and benefit assessment is the population and not the individual patient.

The collider effect arises when the treatment is highly effective and when there are strong unobserved or, for other reasons, not included prognostic factors that could influence time-to-event endpoints. In this situation, the HR is estimated from groups that increasingly differ in their (unobserved or omitted) baseline characteristics over time. However, Strobel et al. (2023) assessed the suspected collider effect in 27 large RCTs and found no empirical evidence of a collider effect. In our view, this observation can be interpreted as follows. Although a collider effect can occur in theory in situations with strong heterogeneity, this is of limited practical relevance because strong heterogeneity usually does not occur in RCTs. Adequately conducted RCTs commonly have clear inclusion and exclusion criteria, leading to sufficient homogeneity, which avoids large biases from potential collider effects.

In summary, we fully agree with Beyersmann et al. (2025) that hazards constitute key quantities for time-to-event data and effect measures such as the HR from Cox proportional hazards models have a valid causal interpretation. In addition, the HR is also a useful population-level estimand with a valid causal interpretation in RCTs with parallel group design. Situations, which could be problematic in theory due to strong heterogeneity, do not occur in typical RCTs. Therefore, the HR can be causally interpreted in RCTs, and the suspected collider effect is usually of limited practical relevance without consequence for decision-making.

Acknowledgments

We thank the editors of the *Biometrical Journal* for the opportunity to comment on this useful paper by Jan, Claudia, and Martin.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The authors have nothing to report.

References

Bender, R., L. Beckmann, and S. Lange. 2016. "Biometrical Issues in the Analysis of Adverse Events Within the Benefit Assessment of Drugs." *Pharmaceutical Statistics* 15, no. 4: 292–296. https://doi.org/10.1002/pst.1740.

Beyersmann, J., C. Schmoor, and M. Schumacher. 2025. "Hazards Constitute Key Quantities for Analyzing, Interpreting and Understanding Time-to-Event Data." *Biometrical Journal* 67: e70057. https://doi.org/10.1002/bimj.70057.

Fay, M. P., and F. Li. 2024. "Causal Interpretation of the Hazard Ratio in Randomized Clinical Trials." *Clinical Trials* 21, no. 5: 623–635. https://doi.org/10.1177/17407745241243308.

Strobel, A., A. Wienke, and O. Kuss. 2023. "How Hazardous Are Hazard Ratios? An Empirical Investigation of Individual Patient Data from 27 Large Randomized Clinical Trials." *European Journal of Epidemiology* 38, no. 8: 859–867. https://doi.org/10.1007/s10654-023-01026-z.