Biometrics, 2026, 00(0), ujag037
https://doi.org/10.1093/biomtc/ujag037
Published: 10 March 2026
Reader Reaction

Comment on "Double robust conditional independence test for novel biomarkers given established risk factors with survival data"

Lucas Kook *

Institute for Statistics and Mathematics, Vienna University of Economics and Business, 1020 Vienna, Austria

*Corresponding author: Lucas Kook, Institute for Statistics and Mathematics, Vienna University of Economics and Business, 1020 Vienna, Austria (lucasheinrich.kook@gmail.com).

Abstract

In their paper, Yang et al. tackle the important challenge of identifying biomarkers that are predictive for a time-to-event response while taking into account relevant risk factors. The proposed solution is a doubly robust conditional independence test based on the model-X framework, that is, the test relies on sampling from the distribution of the biomarker of interest conditional on the relevant risk factors. However, the paper falls short of helping biometricians to make an informed decision on when the proposed test can be used and what alternative doubly robust tests exist in the literature with directly usable and open source software implementations. This comment intends to close this gap by (i) discussing the assumptions on the censoring mechanism that are sufficient for the test to be valid; and (ii) providing a small scale empirical comparison between the test by Yang et al. and another established doubly robust conditional independence tests for time-to-event responses based on the Generalised Covariance Measure. The results show that the test by Yang et al. performs on par with the existing test in terms of type I error control and power, while being computationally more expensive due to the refitting steps. Code to reproduce all results is available in the Supplementary Materials and on GitHub.

Keywords: conditional independence testing, doubly robust statistical inference, Generalised Covariance Measure, survival analysis

1 Introduction

With the more frequent collection of high-throughput experiment data, discovering which of the measured variables are potential biomarkers for a time-to-event response has become an important challenge. Conditional independence tests can address this challenge while being able to account for known risk factors. Yang et al. (2025) propose one such test and we follow their notation to formalize the problem.

Let $T \in \mathbb{R}_+$ denote the survival time, $C \in \mathbb{R}_+$ denote the censoring time, $X \in \mathbb{R}^{d_X}$ denote biomarkers, and $Z \in \mathbb{R}^{d_Z}$ denote other established risk factors for T. We assume that (T, C, X, Z) follows a joint distribution P and that we have access to n observations that are independent and identically distributed copies of $(Y := \min(T, C), \Delta := \mathbb{1}(T < C), X, Z)$.

To establish whether a novel (set of) biomarker(s) is predictive of the event time conditional on established risk factors, we are interested in testing the null hypothesis,

$$H_0 : T \perp\!\!\!\perp X \mid Z, \tag{1}$$

which involves the true but only partially observed survival time T.

We now turn to the assumptions employed by Yang et al. (2025). For our discussion, it will be convenient to split Assumption 1 (i) and (ii) in Yang et al. (2025) into two.

Assumption 1 (Uninformative censoring): It holds that $T \perp\!\!\!\perp C \mid X, Z$.

Assumption 1 amounts to the classical assumption of uninformative censoring, which, as Yang et al. (2025) discuss, is fulfilled if censoring is purely administrative. However, Yang et al. (2025) make a second assumption (Assumption 2 below) which we will show to be unnecessary for the validity of a test for H_0 in (1).

Assumption 2 (Biomarker-independent censoring): It holds that $C \perp\!\!\!\perp X \mid Z$.

Assumption 2 rules out associations between the biomarkers X and the censoring time C other than through the established risk factors Z. Again, if censoring is purely admin-

Received: 5 November 2025. **Accepted:** 12 November 2025

istrative, this assumption is met. Neither Assumption 1 nor Assumption 2 are testable without making further assumptions, as they involve partially unobserved event or censoring times, and their validity, therefore, needs to be argued based on domain knowledge.

2 Discussion of assumption 2

Yang et al. (2025) motivate Assumption 2 with "[r]isk factors Z (patient's characteristics and clinical variables) affect T and C, while biomarkers impact T but not C when conditioned on Z." However, this assumption is not necessary for the validity of the classical likelihood ratio test (Klein et al., 2014) or doubly robust tests of conditional independence for time-to-event responses: There are doubly robust tests of conditional independence (see Section 3) that are valid under Assumption 1, without invoking Assumption 2. We also illustrate empirically that the test by Yang et al. (2025) retains type I error control under violations of Assumption 2 in Section 3.

In Proposition 1, we instead show that Assumption 2 (together with Assumption 1) is sufficient to warrant the use of tests that ignore censoring or use only the binary event indicator to test H_0 in (1). This shows that Assumption 2 is rather too strong, in the sense that one no longer needs tools from survival analysis to test the hypothesis of interest.

Proposition 1: Under H_0 in (1) and Assumptions 1 and 2, the following conditional independencies hold:

$$Y \perp\!\!\!\perp X \mid Z \quad \text{and} \quad \Delta \perp\!\!\!\perp X \mid Z. \tag{2}$$

Proof. Together, Assumptions 1 and 2 imply $(T, X) \perp\!\!\!\perp C \mid Z$, by the conditional contraction property of conditional independence (Dawid, 1979). By conditional weak union (Dawid, 1979), we have $X \perp\!\!\!\perp C \mid T, Z$, which together with H_0, again by conditional contraction, yields $X \perp\!\!\!\perp (C, T) \mid Z$. Now, since $Y = \min(T, C)$ and $\Delta = \mathbb{1}(T < C)$ are both measurable functions of T and C, we obtain (2), which completes the proof. □

The conditional independencies in (2) contain fully observed quantities only and, therefore, allow the use of conditional independence tests that are oblivious to censoring or, as often done, solely consider the event indicator as a binary outcome. In Section 3, we show empirically that conditional independence tests using only the event indicator (but not time and censoring information jointly) are quite sensitive to violations of Assumption 2.

3 Comparison with the tram-GCM test

The goal of this section is to empirically compare the proposed test by Yang et al. (2025) with other existing (doubly robust) conditional independence test in light of the discussion of the assumptions in Section 2, the complexity of the survival and biomarker regression, and computational costs.

Kook et al. (2025) propose the Transformation Model Generalised Covariance Measure (TRAM-GCM) test, which relies on testing an implication of H_0 under Assumption 1, namely,

$$T \perp\!\!\!\perp X \mid Z \Rightarrow \mathbb{E}[(\Delta - \Lambda(Y \mid Z))(X - \mathbb{E}[X \mid Z])] = 0, \tag{3}$$

where $\Lambda(y \mid z)$ denotes the cumulative hazard function of Y given Z, and $\Delta - \Lambda(Y \mid Z)$ are martingale residuals (Therneau et al., 1990). The test requires a survival regression to obtain an estimate of Λ and a conditional mean regression to obtain an estimate of $\mathbb{E}[X \mid Z]$, both of which may be based on machine learning fulfilling certain rate conditions (Kook et al., 2025, Theorem 15). The TRAM-GCM test is valid under arbitrary forms of uniformative left, right, and interval censoring (see Kook et al., 2025, supplementary material). Here, the test is presented only for the case of right censoring, in line with the setting in Yang et al. (2025). Model-X based tests (Candès et al., 2018) and the GCM test (Shah and Peters, 2020) for non-censored responses have been compared in detail in Niu et al. (2024). A doubly robust approach to estimating hazard ratios has been proposed in Vansteelandt et al. (2024).

We conduct a small-scale empirical study (similar to setting MX1 and MT2 in Yang et al., 2025) in which T is generated from a Weibull model with scale $\lambda = 0.1$, shape $\rho = 2$, and hazard ratio $\exp(0.8Z + \beta X)$ (hence, H_0 holds iff $\beta = 0$); C follows an exponential distribution with rate $0.1 \cdot \exp(0.5Z + 0.5X)$ (thus, Assumption 1 holds while Assumption 2 is violated); Z follows a standard normal distribution and X is generated according to $X = 0.5Z + 0.25\varepsilon$, with standard normally distributed ε.

We apply the test by Yang et al. (2025) using a Cox and a linear model, the TRAM-GCM test with a Cox and a linear model, and the GCM test by Shah and Peters (2020), which uses the binary event indicator as a response, a random forest for regressing Δ on Z and a linear model for regressing X on Z. The latter test is included to illustrate the danger of using only the event indicator when Assumption 2 is violated.

Figure 1(A) shows the empirical cumulative distribution function (ECDF) of the P-values when sampling 300 observations from the setting described above, repeated 100 times, which, for a valid test and under the null hypothesis ($\beta = 0$), are expected to be standard uniformly distributed. For $\beta = 0$, neither the Yang et al. (2025) test nor the TRAM-GCM test show violations of type I error control at any significance level, despite violations of Assumption 2, while the binary GCM test violates type I error control. These results emphasize empirically that Assumption 2 is not necessary for ensuring type I error control. In terms of power ($\beta = 0.5$), the test by Yang et al. (2025) and TRAM-GCM perform on par in this setup.

In practice, computational time is another important factor to consider when choosing an appropriate test. Figure 1(B) shows that the test by Yang et al. (2025) is computationally expensive, leading to median runtimes that are almost two orders of magnitude slower than the TRAM-GCM test (using the same classes of regression models), which is due to re-fitting of the survival regressions for each model-based bootstrap step. For this comparison we used $B = 199$ resamples, which is rather moderate compared to $B = 1000$ recommended in Yang et al. (2025). In this case, the binary GCM test is slower

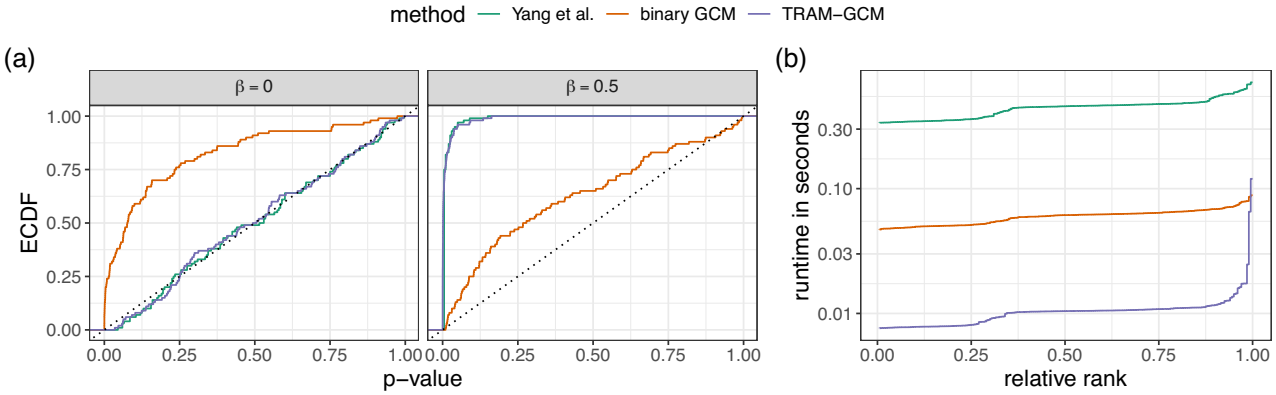

Figure 1 Comparison of the test proposed by Yang et al. (2025), the TRAM-GCM test, and a GCM test that uses the binary event indicator (binary GCM) based on $n = 300$, $d_X = d_Z = 1$, 100 repetitions, and using $B = 199$ bootstrap iterations. In the DGP, Assumption 2 is violated. (A) ECDFs of the 100 computed *P*-values per test under H_0 ($\beta = 0$) and an alternative $\beta = 0.5$. The dotted line indicates a uniform distribution. (B) Empirical quantile function of runtimes for all three tests. ECDF, empirical cumulative distribution function; TRAM-GCM, Transformation Model Generalised Covariance Measure.

than the TRAM-GCM test, as it relies on a random forest regression instead of a Cox model.

We additionally observe that although Yang et al. (2025) allow for $d_X > 1$ in their theoretical results and algorithms, their implementation is currently limited to $d_X = 1$ and linear regression for the X on Z regression. The TRAM-GCM test is implemented in the `R` package **comets** (Kook and Lundborg, 2024) and supports parametric, Cox, and Random Forest-based survival regression and a multitude of algorithms for the X on Z regression, including XGboost and random forests, and allows for X with $d_X > 1$ of discrete, ordinal, continuous, or mixed type.

4 Conclusion

Yang et al. (2025) propose a powerful doubly robust conditional independence test under the model-X framework. We critically discuss which assumptions are sufficient for the validity of the proposed and related conditional independence tests and provide a small scale empirical comparison with the TRAM-GCM test. Biomarker-independent censoring (Assumption 2) is *not* necessary for the validity of the test by Yang et al. (2025) or the TRAM-GCM test, while it is sufficient for the validity of tests ignoring censoring or using the binary event indicator as an outcome. Empirically, both tests perform similarly under the considered data generating mechanism in terms of type I error control and power, while the TRAM-GCM test is computationally faster. Violations of Assumption 2 can have a detrimental impact on type I error control of tests using only the event indicator.

Both the discussion and the empirical comparison, for which code is available in the Supplementary Materials and on GitHub, are intended to inform biometricians and practitioners in their choice of test for modern feature selection and causal inference problems in applications with time-to-event responses. Future work is needed for a more nuanced empirical comparison of the available tests and related methods under broader data generating processes and model misspecification, requiring a more flexible implementation of the test in Yang et al. (2025). More generally, doubly robust conditional independence tests that are valid under dependent censoring are an interesting avenue for future research.

Acknowledgments

I thank Robert Bajons for fruitful discussions.

Supplementary materials

Supplementary material is available at *Biometrics* online.

Code is available with this paper at the Biometrics website on Oxford Academic and on GitHub at https://github.com/LucasKook/doubly-robust-survival

Funding

None declared.

Conflicts of interest

None declared.

Data availability

No new data were generated or analyzed in support of this research.

References

Candès, E., Fan, Y., Janson, L. and Lv, J. (2018). Panning for gold: 'Model-X' knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80, 551–577.

Dawid, A. P. (1979). Conditional independence in statistical theory. *Journal of the Royal Statistical Society B*, 41, 1–15.

Klein, J. P., Van Houwelingen, H. C., Ibrahim, J. G. and Scheike, T. H. (Eds.). (2014). *Handbook of Survival Analysis*(1st ed.). Boca Raton, FL: CRC Press.

Kook, L. and Lundborg, A. R. (2024). Algorithm-agnostic significance testing in supervised learning with multimodal data. *Briefings in Bioinformatics*, 25, bbae475.

Kook, L., Saengkyongam, S., Lundborg, A. R., Hothorn, T. and Peters, J. (2025). Model-based causal feature selection for general response types. *Journal of the American Statistical Association*, 120, 1090–1101.

Niu, Z., Chakraborty, A., Dukes, O. and Katsevich, E. (2024). Reconciling model-X and doubly robust approaches to conditional independence testing. *The Annals of Statistics*, 52, 895–921.

Shah, R. D. and Peters, J. (2020). The hardness of conditional independence testing and the Generalised Covariance Measure. *The Annals of Statistics*, 48, 1514–1538.

Therneau, T. M., Grambsch, P. M. and Fleming, T. R. (1990). Martingale-based residuals for survival models. *Biometrika*, 77, 147–160.

Vansteelandt, S., Dukes, O., Lancker, K. V. and Martinussen, T. (2024). Assumption-lean Cox regression. *Journal of the American Statistical Association*, 119, 475–484.

Yang, B., Qin, J., Ning, J. and Liu, Y. (2025). Double robust conditional independence test for novel biomarkers given established risk factors with survival data. *Biometrics*, 81, ujaf133.

Received: 5 November 2025. **Accepted:** 12 November 2025