Received: 25 March 2022 | Accepted: 1 August 2022

DOI: 10.1111/biom.13794

Biometrics WILEY

DISCUSSION

Discussion on "Instrumental variable estimation of the causal hazard ratio" by Linbo Wang, Eric Tchetgen Tchetgen, Torben Martinussen, and Stijn Vansteelandt

A. James O'Malley[1,2] | **Pablo Martínez-Camblor**[1,3] | **Todd A. MacKenzie**[1,2]

[1]Department of Biomedical Data Science, Geisel School of Medicine at Dartmouth, Hanover, New Hampshire, USA

[2]The Dartmouth Institute for Health Policy and Clinical Practice, Geisel School of Medicine at Dartmouth, Hanover, New Hampshire, USA

[3]Department of Anesthesiology, Dartmouth Health, Lebanon, New Hampshire, USA

Correspondence
A. James O'Malley, Department of Biomedical Data Science and The Dartmouth Institute for Health Policy and Clinical Practice, Geisel School of Medicine at Dartmouth, Hanover, NH 03755, USA.
Email: James.OMalley@dartmouth.edu

KEYWORDS
causal hazard ratio, conditional causal model, instrumental variable, marginal causal model, principal stratification weights, unmeasured confounder–instrumental variable interaction

1 | INTRODUCTION

We congratulate Professors Wang, Tchetgen Tchetgen, Martinussen, and Vansteelandt for their research leading to this intriguing paper, which has generated a lot of discussion among ourselves and that we expect will also generate a lively discussion among the statistics community at large. The combination of unmeasured confounding and censored outcomes together with the use of the semiparametric Cox model is a challenging problem. The quest for estimators with appealing statistical properties for estimating causal quantities such as the causal hazard ratio has recently received considerable attention and seen the development of several distinct approaches. Therefore, the development of Wang et al. (2022) and the opportunity to discuss it is timely. In the following, we focus on the reasonableness of an assumption underlying the causal hazard ratio estimator, characterize the causal hazard ratio estimator by demonstrating its correspondence to an existing estimator not mentioned in their paper, consider issues related to the assumed data-generating mechanism, and present additional simulations to present a fairer comparison of the Wang et al. (2022) estimator to an existing estimator than presented in their paper.

2 | REASONABLENESS OF ASSUMPTION A5

The key variables are the possibly right-censored continuous survival outcome T, the treatment D, a vector of observed confounders X, an unobserved confounder U, and an instrumental variable Z. The causal model assumed in Wang et al. (2022) is given by the marginal structural model

$$\lambda_d^T(t) = \lambda_0^T(t)\exp(\psi d), \tag{1}$$

where $\lambda_d^T(t) = -[S_d^T(t)]'/S_d^T(t)$ is the potential hazard function and $S_d^T(t)$ the potential survival function at follow-up time t. The target parameter is ψ, the log causal hazard ratio. The standard instrumental variable assumptions (A1–A3), the independent censoring assumption (A4), and a no additive $U - Z$ interaction assumption (A5) are relied upon to identify a consistent estimator of ψ.

We have doubts about the appropriateness of A5 in Wang et al. (2022), which we restate:

$$\begin{aligned}\delta^D(X,U) &= E[D \mid Z=1,X,U] - E[D \mid Z=0,X,U] \\ &= E[D \mid Z=1,X] - E[D \mid Z=0,X] \\ &= \delta^D(X),\end{aligned} \tag{2}$$

As noted in Remark 3 of Wang et al. (2022), Equation (2) implies that the treatment selection mechanism cannot follow the logistic or other nonlinear form, except in the case that $E[D \mid Z, X, U]$ is independent of U, which would imply that U is not a confounder. (Tangentially, we note that in the highly simplified case in which there are no measured covariates ($X = \emptyset$) and U is binary, the logistic regression model could satisfy Equation (2) as one can solve for the model parameters so that $E[D \mid Z = 1, U = 1] - E[D \mid Z = 0, U = 1] = E[D \mid Z = 1, U = 0] - E[D \mid Z = 0, U = 0]$. But due to the noncollapsibility of the logistic regression model, a solution will not hold in any generality.) Therefore, the requirement in Equation (2) appears unnatural and restrictive. Given that D is binary, $E[D \mid Z = 1, X, U]$ is a probability and so one would naturally expect a change in X or U to result in a smaller change if the baseline probability was near 0 or 1 compared to near 0.5. In the generalized linear models family, the link function typically allows equivalent changes on an unrestricted scale to be mapped to the scale of the outcome, where the resulting fitted value is guaranteed to be within range and the effect-size shrinks near the end of the range (0 and 1). By assuming no interaction between X and U on the probability scale, A5 induces an interaction effect on the log-odds or other link-transformed scale such that a larger effect is implied when the baseline probability of treatment is close to 0 versus close to 0.5. Furthermore, if an instrument Z is assumed to have a constant effect (say $E[D \mid Z = 1, X, U] - E[D \mid Z = 0, X, U] = 0.1$) on treatment selection, then what happens if the baseline expectation (probability) with respect to Z, $E[D \mid Z = 0, X = x, U = u]$, is 0.95? It appears that A5 is made for theoretical convenience. However, in a real data situation, we think that the reasonableness of A5 deserves more consideration.

For example, one might ask: What are the implications of a violation of A5? How robust are the findings to violations of A5, such as alternative scenarios in which the treatment selection mechanism follows the logistic or probit form given (Z, X, U)?

3 | A HAZARD RATIO IN THE COMPLIERS?

In Wang et al. (2022), the inclusion of measured covariates is accommodated in their time-to-event model for binary treatments by using inverse weighting by the propensity of the binary instrument; that is, $f(Z \mid X)$. If there are no measured covariates the weight function they employ, $w_0(Z, X) = (2Z - 1)/\{f(Z|X)\delta^D(X)\}$, is proportional to $w_0(Z) = (2Z - 1)/f(Z)$. To avoid a numerator in their estimating equation having an expected value of zero they add a *stabilization term* that depends on the treatment, $h(D)$, yielding the weight function $w(Z, D) = (2Z - 1)h(D)/f(Z)$ for which it is asserted that $h(0)$ and $h(1)$ should have opposite signs. The probability masses, $f(0)$ and $f(1)$, for the distribution of Z can be estimated by $n_{0\cdot}/n$ and $n_{1\cdot}/n$, where n is the sample size and $n_{z\cdot} = n_{z0} + n_{z1}$, with n_{zd} the number of subjects for which $Z = z, D = d$. Table 1 shows the weights of Wang et al. (2022) according to the values of the binary instrument and binary treatment. An interesting aspect of Wang et al. (2022) is the use of weights that take negative values, which is usually considered to be an error or ignored in most statistical software (e.g., R, Stata, or SAS).

TABLE 1 The weights of Wang et al (2022) and the *principal stratification weights* of MacKenzie et al (2016) by binary instrument and binary exposure, where $m = n_{00}n_{11} - n_{01}n_{10}$

Instrument	Treatment	Wang et al (2022) weights	Principal stratification weights
$Z = 0$	$D = 0$	$-h(0)n/n_{0\cdot}$	$+m^{-1}c_0 n_{0\cdot} n_{1\cdot}$
$Z = 0$	$D = 1$	$-h(1)n/n_{0\cdot}$	$-m^{-1}c_1 n_{1\cdot}^2$
$Z = 1$	$D = 0$	$+h(0)n/n_{1\cdot}$	$-m^{-1}c_0 n_{0\cdot}^2$
$Z = 1$	$D = 1$	$+h(1)n/n_{1\cdot}$	$+m^{-1}c_1 n_{0\cdot} n_{1\cdot}$

The right-most column in Table 1 shows the weights proposed by MacKenzie et al. (2016) for estimating the hazard ratio in the compliers in a randomized clinical trial in which patients could be compliers, always-takers, or never-takers (the most general situation that is consistent with the assumption of monotonicity, which rules out defiers). That approach applied principal stratification (Frangakis & Rubin, 2002) to the empirical distribution of right-censored times to events and also yields weights that could be negative. The terms c_0 and c_1 are arbitrary positive constants, and the proposed principal stratification estimator of the hazard ratio is invariant to these two constants.

We were pleasantly surprised to determine that if we choose $h(0) < 0$ (in which case $h(1) > 0$) and take

$$c_0 = -\frac{h(0)nm}{n_{0\cdot}^2 n_{1\cdot}} \text{ and } c_1 = \frac{h(1)nm}{n_{0\cdot} n_{1\cdot}^2}, \tag{3}$$

where $m = n_{00}n_{11} - n_{01}n_{10}$, for the constants in our *principal stratification weights* that we obtain the same estimating equation as in Wang et al. (2022). Therefore, we propose that in the case of no measured covariates, the estimator of the marginal hazard ratio ψ is an estimator of the marginal hazard ratio in the compliers. We also note that the estimator in MacKenzie et al. (2016) may be converted to an estimator of the population-averaged hazard ratio; this can be shown by exploiting the facts that $Z \perp (T, C) \mid$

D, X, U (the exclusion restriction, which states that the potential survival and censoring times for always-takers and never-takers are invariant to Z conditional on D, X, U) and that the proportion of compliers in the population is estimable; for example, by the method-of-moments estimator $\Pr(Complier) = m/(n_{0\cdot}n_{1\cdot})$.

4 | CONSIDERATIONS REGARDING DATA-GENERATING MECHANISM

The choice to model the relationship of treatment D to the survival time T marginally is central to Wang et al. (2022). We believe that the assumed data-generating mechanism under which the marginal Cox model $S_d^T(t)$ obtains is not given as much emphasis as it deserves. Although the marginal model depends only on D, the assumptions under which the estimator proposed in Wang et al. (2022) is consistent involve expressions that condition on X and U (e.g., as in Equation 2). From Section S8 of the Supporting Information, we obtain the expression:

$$S_d^T(t) = \int_{x,u} S_{T|D,X,U}(t \mid D = d, X = x, U = u) dF_{X,U}(x,u), \tag{4}$$

where $S_{T|D,X,U}(t \mid D = d, X = x, U = u)$ denotes the conditional causal survival model and $\pi_{X,U}(x,u) = d[F_{X,U}(x,u)]/dxdu$ is the distribution describing the relative frequency of individuals in the population of interest. We find Equation (4) helpful as it clarifies that the mixing distribution $\pi_{X,U}(x,u)$ is not conditional on $D = d$, as would be implied by direct application of the theorem of total probability to $S_{T|D}(t \mid D = d)$, but rather averages $S_{T|D,X,U}(t \mid D = d, X = x, U = u)$ over (X, U) for all values of d, as for the g-formula (Robins, 1986). This also suggests that the contrast of potential outcomes for the marginal causal model $S_d^T(t)$ corresponds to changing D from 0 to 1 for all members of the population characterized by $\pi_{X,U}(x,u)$, another point not entirely clear in Wang et al. (2022). Further, the expansion in Equation (4) justifies use of the phrase *population-average causal model* (Wang et al., 2022; Martínez-Camblor et al., 2022).

In a practical situation, it might be debated whether it is better to have a simple expression for the marginal model versus the conditional model. This consideration has arisen in noncausal situations. For example, in modeling a mixed (part binary, part continuous) outcome in which the goal is to study the relationship of the predictors to the marginal mean of the outcome (*marginal covariate effects*), should the binary and the continuous models be analyzed assuming a marginalized two-part model (MTP) that parameterizes covariate effects directly in terms of the marginal mean of the outcomes (Smith et al. 2014; Smith, Neelon, & Maciejewski 2017; Smith, Neelon, Maciejewski, & Preisser 2017) or conditionally? Under the standard two-part model, the effects of the predictors on the marginal mean have a complicated expression that involves the values of all other predictors included in the model. In contrast, a simple expression is provided under the MTP. In the case of the two-part model, a decision between the two specifications can be informed using model comparison procedures. When $S_{T|D,X,U}(t \mid D = d, X = x, U = u)$ is able to be determined such as for the simulation experiment in Wang et al. (2022), this provides the potential for providing an additional lens through which to justify (or not) the model for $S_d^T(t)$. The general question of whether a marginal causal model, $S_d^T(t)$, versus a conditional causal model, $S_d^T(t \mid X = x, U = u) = S_{T|D,X,U}(t \mid D = d, X = x, U = u)$, is more justified in a given situation is an interesting topic for future research and discussion.

5 | FAIRER COMPARISON TO AN EXISTING CAUSAL HAZARD RATIO ESTIMATOR

We are curious as to why the range of simulation conditions considered in Wang et al. (2022) is so limited in terms of the amount of unmeasured confounding considered, as evinced by the fact that the Cox-crude procedure had at most 5.7% bias under their simulations (see Table 1 of Wang et al. 2022). Furthermore, we believe that the poorer performance of the method in MacKenzie et al. (2014), hereafter the "MacKenzie method," is simply because the published version is for the case in which there are no observed covariates (whether or not these are confounders) and the simulations in Wang et al. (2022) do not consider this case. Therefore, the results in Wang et al. (2022) give the misleading impression that the MacKenzie method is generally flawed when in fact it is simply that it was not designed to be used with covariates. We adapted the simulation study performed in Martínez-Camblor et al. (2022) to address both of these points.

To compare the Wang et al. (2022) and MacKenzie method estimators in the case of no observed covariates, we simulated results for a series of scenarios defined by different values of the treatment effect (the log of the causal hazard ratio ψ), the censoring proportion, and the strength of the instrumental variable. For each simulated dataset, we generated the baseline potential survival times

$$T_{i,0} = -\log\left(1 - F_{\text{Gamma},4,1}(U_i + W_i)\right), \tag{5}$$

TABLE 2 Performance of the Wang et al (2022) and MacKenzie methods for estimating the log causal hazard ratio ψ based on 2000 simulated datasets each with sample size 1000 for different values of e^{ψ}, the censoring percentage, and the instrument strength ($\delta^D = \mathbb{E}[D \mid Z = 1] - \mathbb{E}[D \mid Z = 0]$). Results are reported as bias times 100 (standard error times 100, in parenthesis) and coverage.

			Bias×100 (SE×100)			Coverage rate	
e^{ψ}	Censoring %	δ^D	Cox	Wang	MacKenzie	Wang	MacKenzie
1.0	50	0.41	−47.1 (9.2)	0.79 (26.5)	0.29 (24.3)	0.947	0.957
1.5	50	0.41	−46.3 (9.0)	2.83 (26.6)	2.30 (23.8)	0.952	0.960
2.0	50	0.41	−46.2 (9.2)	4.23 (27.7)	5.68 (26.7)	0.963	0.954
1.0	75	0.41	−55.9 (12.9)	2.46 (37.9)	0.99 (33.7)	0.969	0.960
1.5	75	0.41	−55.9 (14.1)	6.63 (41.0)	3.76 (34.3)	0.971	0.966
2.0	75	0.41	−55.3 (13.9)	8.15 (45.9)	7.56 (37.7)	0.964	0.963
1.0	50	0.30	−70.7 (9.1)	2.58 (40.0)	−0.11 (34.6)	0.951	0.969
1.5	50	0.30	−69.3 (9.2)	8.19 (43.0)	3.34 (35.6)	0.956	0.967
2.0	50	0.30	−69.4 (9.3)	12.3 (52.9)	9.55 (40.9)	0.963	0.961
1.0	75	0.30	−90.6 (14.4)	9.78 (74.8)	0.93 (47.5)	0.958	0.983
1.5	75	0.30	−88.6 (14.3)	14.2 (72.1)	3.50 (50.0)	0.956	0.981
2.0	75	0.30	−88.2 (13.8)	19.2 (76.6)	12.1 (58.7)	0.954	0.986

and assigned an observed survival time $T_i = T_{i,0}(e^{-\psi}D_i + 1 - D_i)$ (hence, $T_{i,1} = T_{i,0}e^{-\psi}$ is the potential survival time under treatment) according to the treatment selection model

$$D_i = I(\alpha_U(U_i - 1) + \alpha_Z Z_i^b + \epsilon_i < 0), \tag{6}$$

where $F_{\text{Gamma},4,1}(.)$ denotes the cumulative distribution function of a gamma distribution with shape 4 and scale 1 that arises from the sum of an independent gamma(1,1) random variable U_i (representing an unmeasured confounder) and an independent gamma(3,1) random variable W_i (representing random factors affecting the survival times), $Z_i^b = I(Z_i > 0)$ is the binary instrumental variable determined from the independent standard normal random variable Z_i (representing the underling propensity of the instrument), and ϵ_i (representing random factors affecting treatment selection) is an independent standard normal random variable. In addition, the observed censoring times are given by $C_i = C_{i,0}(e^{-\psi}D_i + 1 - D_i)$, where the baseline potential censoring time $C_{i,0}$ is an independent random variable from an exponential distribution with mean 1 (i.e., a gamma(1,1) distribution). The effect of treatment on the censoring times is thus the same as its effect (ψ) on the survival times, whereas the unmeasured confounder has no effect on the censoring times. For each scenario, we performed 2000 Monte Carlo runs of hypothetical studies each comprising 1000 patients and computed the bias, average standard error, and coverage of the 95% confidence interval for the target parameter ψ. We also include the bias for the regular Cox model as a measure of the impact of the unobserved covariate (and sometimes confounder) in the conditional hazard.

The simulation results in Table 2 reveal that the Wang et al. (2022) and MacKenzie methods both experience only modest amounts of bias and obtain coverage above the nominal level in almost all instances. It is also clear that these settings embody a substantial amount of unmeasured confounding with the Cox-crude results having up to nearly 100% bias. While the MacKenzie method had lower bias in the majority of scenarios, it did not on every occasion and we have only considered a small number of scenarios even within the no observed covariates case, leading to our conjecture that neither the Wang et al. (2022) nor the MacKenzie method makes the other inadmissible.

6 | CONCLUSION

We again congratulate the authors on their thought provoking and excellent paper, which we thoroughly enjoyed reading and discussing. We look forward to future robust conversations on this and related topics!

ORCID

A. James O'Malley https://orcid.org/0000-0001-8389-6217
Todd A. MacKenzie https://orcid.org/0000-0002-0215-2003

REFERENCES

Frangakis, C.E. & Rubin, D.B. (2002) Principal stratification in causal inference. *Biometrics*, 58(1), 21–29.
MacKenzie, T., Tosteson, T., Morden, N., Stukel, T. & O'Malley, A.J. (2014) Using instrumental variables to estimate a Cox's pro-

portional hazards regression subject to additive confounding. *Health Services and Outcomes Research Methodology*, 14(1), 54–68.

MacKenzie, T.A., Løberg, M. & O'Malley, A.J. (2016) Patient centered hazard ratio estimation using principal stratification weights: application to the norccap randomized trial of colorectal cancer screening. *Observational Studies*, 2, 29–50.

Martínez-Camblor, P., MacKenzie, T.A. & O'Malley, A.J. (2022) Estimating population-averaged hazard ratios in the presence of unmeasured confounding. *International Journal of Biostatistics*, doi: 10.1515/ijb–2021–0096.

Robins, J.M. (1986) A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7, 1393–1512.

Smith, V., Neelon, B. & Maciejewski, M. (2017) A marginalized two-part model for longitudinal semicontinuous data. *Statistical Methods in Medical Research*, 26(4), 1949–1968.

Smith, V., Neelon, B., Maciejewski, M. & Preisser, J.S. (2017) Two parts are better than one: modeling marginal means of semicontinuous data. *Health Services and Outcomes Research Methodology*, 17, 198–218.

Smith, V., Preisser, J.S., Neelon, B. & Maciejewski, M. (2014) A marginalized two-part model for semicontinuous data. *Statistics in Medicine*, 33, 4891–4903.

Wang, L., Tchetgen Tchetgen, E., Martinussen, T. & Vansteelandt, S. (2022) Instrumental variable estimation of the causal hazard ratio. *Biometrics*, In press. https://doi.org/10.1111/biom.13792

How to cite this article: O'Malley, A.J., Martínez-Camblor, P. & MacKenzie, T.A. (2023) Discussion on "Instrumental variable estimation of the causal hazard ratio" by Linbo Wang, Eric Tchetgen Tchetgen, Torben Martinussen, and Stijn Vansteelandt. *Biometrics*, 79, 559–563. https://doi.org/10.1111/biom.13794