Political Science Research and Methods (2025), 1–17
doi:10.1017/psrm.2025.12

ORIGINAL ARTICLE

Out of order: specification check sequencing in Cox models

Benjamin T. Jones[1] and Shawna K. Metzger[2]

[1]Department of Political Science, University of Mississippi, Oxford, MS, USA and [2]Department of Political Science, University at Buffalo, Buffalo, NY, USA
Corresponding author: Shawna K. Metzger; Email: smetzger@buffalo.edu

(Received 1 August 2023; revised 12 August 2024; accepted 12 October 2024)

Abstract

The Cox duration model serves as the basis for more complex duration models like competing risks, repeated events, and multistate models. These models make a number of assumptions, many of which can be assessed empirically, sometimes for substantive ends. We use Monte Carlo simulations to show the order in which practitioners assess these assumptions can impact the model's final specification, and ultimately, can produce misleading inferences. We focus on three assumptions regarding model specification decisions: proportional hazards (PH), stratified baseline hazards, and stratum-specific covariate effects. Our results suggest checking the PH assumption before checking for stratum-specific covariate effects tends to produce the correct final specification most frequently. We reexamine a recent study of the timing of GATT/WTO applications to illustrate our points.

Keywords: duration models; survival analysis; Cox models; proportional hazards assumption; assumption tests

Political scientists use duration models to investigate hypotheses about an event's occurrence. When the event of interest is part of a larger process, we can use more complex duration models, such as competing risks (e.g., Conrad and Moore, 2010; Wong and Chun-Man Chan, 2021), repeated events (e.g., Blair et al., 2022), or multistate models (e.g., Jones and Metzger, 2018; Carter and Lemke, 2022), to utilize information about all the process' constituent parts as a partial guard against certain kinds of incorrect inference. However, as the complexity of duration models grows, the number of specification decisions and other assumptions that researchers should check do as well. Our interest is in whether the order in which we evaluate these decisions matters, given practitioners' general aim of arriving at the correct model specification.

Three specification decisions are particularly important. First, from the simplest duration models, we know proportional-hazard models like the exponential, Weibull, Gompertz, and Cox assume covariate effects are proportional with time—the *proportional hazards (PH) assumption*. Second, in any duration model, we can permit different groups to experience our event of interest at different underlying rates using a *stratified baseline hazard*, which estimates a unique baseline hazard for each group (stratum), amounting to group/stratum-specific intercepts that vary at each *t* value. Stratifying can be used for a number of reasons, such as addressing a PH assumption violation or more substantively motivated reasons, with "groups" potentially ranging from literal groups to certain event types. Third, in yet another batch of related models, we can also permit our covariates to have different effects for different groups or for different event types, if the process of interest comprises multiple events that subjects can experience (*stratum-specific covariate effects*, also known as transition-specific covariate effects).

Different segments of the duration modeling literature are synonymous with different combinations of these assumption-driven specification decisions; we refer to these three specification decisions as "assumptions" for succinctness hereafter. All PH models are characterized by the PH assumption and researchers are often trained to check for violations when first learning about these models. Repeated events models frequently discuss stratifying on number of prior events, but not stratum-specific covariate effects. Competing risks models employ both stratified baseline hazards and stratum-specific covariate effects in their starting specifications, as do multistate models. Importantly, repeated events and competing risks models are special cases of multistate models, underscoring that all three assumptions are potentially relevant for these models, regardless of the frequency with which they are traditionally discussed in those literatures.

Given all three assumptions are in play for all complex duration models, in what order should we check and implement corrections for each assumption? In general, scholars are paying increasing attention to the garden of forking paths—the idea that the order in which we make research design decisions can unintentionally affect our inferences' veracity (Gelman and Loken, 2014). Within the duration modeling literature, the potential interconnectivity among testing the three assumptions we note has received little attention, even though reasons exist to suspect the conclusions we draw from one test may be contingent on which assumptions we have already checked.

We use Monte Carlo simulations to provide evidence-based guidance about the order in which to check these three assumptions for Cox duration models. We craft our simulations to represent a best-case scenario: a large *n*, no right censoring, two possible strata, and two uncorrelated covariates. Our results thus demonstrate which sequence of tests is most likely to generate a correctly specified model under optimal conditions. Our initial Cox specification follows the standard multistate model specification by assuming both different baseline hazards and different covariate effects across strata. We consider six different testing sequences that researchers might employ and use our starting specification to assess each sequence's performance.

Our simulations show the crux of the matter is the PH assumption, and ensuring researchers do not induce or mask PH violations through the decisions they make about covariate effects. In our scenarios, we find researchers are generally better off testing for stratum-specific baseline hazards first, then for PH violations (and correcting for any violations), then testing whether stratum-specific covariate effects are appropriate. The key is testing and correcting for PH violations before testing for stratum-specific covariate effects. Otherwise, PH violations may give the appearance that a covariate's effect is collapsible across strata when it is not, in truth, leading us to an incorrect specification decision that can affect any remaining assumption checks.

Our simulation results imply that order does matter, even in the best-case scenarios we investigate. It is reasonable to suspect the same will hold for many applied situations, as the characteristics of such data generally work against drawing the correct conclusions from our tests of interest (e.g., fewer subjects, correlated covariates, presence of right censoring). Knowing that order matters is therefore important information for practitioners who have substantive hypotheses about groups having different baseline hazards, groups having different covariate effects, and/or groups having different time-varying covariate effects. Our reexamination of Davis and Wilf's (2017) study of time to apply for GATT/WTO membership highlights one such situation; Maeda (2010) is potentially another. We find Davis and Wilf's key geopolitical covariates do impact countries' GATT/WTO application rates, in line with their main argument. However, we find this effect exists only for former colonies, which is a new insight.

We begin by formally stating the three assumptions we refer to and articulate how our conclusions about each assumption could affect our conclusions about the other two. Next, we explain our

https://doi.org/10.1017/psrm.2025.12 Published online by Cambridge University Press

simulation setup. Third, we discuss our simulation results, focusing on the two orderings we believe researchers are most likely to employ in practice due to ease of implementation. We then rerun Davis and Wilf's analysis to illustrate the implications of our simulation findings for practitioners, where we uncover the scope condition related to their key covariates of interest. Finally, we provide concluding remarks.

1. The specification checks

In a basic proportional-hazard duration model, the hazard of an event occurring equals (Box-Steffensmeier and Zorn, 2001, 974):

$$h(t) = h_0(t)\exp(\beta' X) \tag{1}$$

where t measures how long the subject has been at risk of experiencing the event; $h_0(t)$ is the baseline hazard of the event occurring at t; and β is a vector of coefficient estimates relating the covariates, X, to the hazard of the event occurring. We could also synonymously conceive of this event as a *transition* between two stages: "event not experienced yet" and "event experienced," as certain duration model literatures do. Our focus is on semi-parametric duration models, for tractability, with the Cox being the most well-known. However, our broader point about the potential impact of assumption-check ordering on final coefficient estimates applies to all regression models.

To illustrate these points, we use Davis and Wilf's (2017) (subsequently, "D&W") study about the timing of countries' membership applications to the GATT/WTO as our running example. Their dependent variable is the length of time that passes between when a country becomes eligible to apply for GATT/WTO membership and when the country formally applies. A model of equation 1's form for D&W would estimate the time until a country applies to join the GATT/WTO, the event of interest. We could equivalently express this in terms of the time until a country *transitions* from eligibility to submitting a formal application. Eq. 1's form implies applying to join occurs at the same underlying rate for all countries after conditioning on covariates ($h_0(t)$), and that each covariate in the model has an equivalent effect for all countries (β).

We frame our discussion in terms of (semi-parametric) multistate duration models, which are duration models that permit transitions among multiple risksets ("stages"), including recursive transitions to prior stages (Metzger and Jones, 2016). As Metzger and Jones (2016) discuss, basic duration models, repeated events models, and competing risks models are all special cases of the multistate model; "event type" and "transition" are synonyms. The hallmark of all multistate models is a situation in which one or more groups of subjects is at risk of one or more possible transitions.

We find the multistate framing useful for two reasons. First, the multistate literature regularly discusses evaluating our three assumptions. By contrast, one or more of the three assumptions are conspicuously less discussed in segments of the duration modeling literature corresponding to special cases of the multistate model, despite these assumptions still guiding final model specifications, often implicitly. For instance, repeated events models concern situations in which a subject may be at risk of experiencing the same event multiple times. Typically, repeated events models stratify the baseline hazard by an event counter, capturing the notion that the risk of experiencing a first event across time likely differs from the risk of experiencing a sixth event across time. Conversely, though, repeated events models typically hold covariate effects constant and generally say little about the possibility of relaxing this assumption, which effectively assumes covariate effects do not vary based on the number of prior events. As another example, the competing risks model captures situations where all subjects begin in the same stage, but are simultaneously at risk of two or more transitions. Such models typically use stratified baseline hazards and permit covariate effects to differ across each possible transition, but discussions of whether either choice is necessary, empirically, are rarer.

Second, multistate models offer a useful starting point for an initial model specification upon which to perform the various specification tests. The typical starting point for multistate models,

https://doi.org/10.1017/psrm.2025.12 Published online by Cambridge University Press

as well as competing risks models, is a model that (1) includes stratified baseline hazards on transition and (2) allows covariate effects to vary across transitions/strata. This specification offers a notable advantage for subsequent specification tests, relative to other possibilities. Therneau and Grambsch (2000, 148) note the PH test may return significant results not only in the presence of PH violations, but also other forms of model misspecification.[1] By estimating a starting model that is fully stratified, with distinct covariate effects across strata, researchers can avoid model misspecification stemming from violating those assumptions from the outset (e.g., forcing a covariate's coefficient to be the same across strata, when it may differ in reality).

The potential costs of misspecification are high. If a researcher improperly collapses a covariate's effect, the resulting estimated coefficient may be biased—i.e., the collapsed estimate may differ from the covariate's true effect in each stratum, which could potentially yield erroneous inferences from subsequent PH tests. Missing a PH violation can also induce bias, both for the violating covariate and other covariates in the model, potentially (Keele, 2008, 6).

By contrast, using multistate models' initial specification strategy has lower costs. The initial specification may be somewhat inefficient, particularly in small sample sizes, because of the number of parameters for which the model must obtain estimates. However, the estimates will be unbiased, and we generally err toward avoiding bias in our estimates, as any bias may influence our conclusions about the optimal sequence of specification tests. Furthermore, with our interest in establishing the ideal sequence under the best of circumstances, we deliberately run our simulations with large sample sizes, mooting concerns about inefficiency in our specific context.

1.1. Whether to stratify the baseline hazard

1.1.1. What is it?

The baseline hazard represents the underlying rate at which subjects experience an event across t when all substantive covariates are equal to zero. "To stratify the baseline hazard" means to permit different groups to experience the event of interest at a different underlying rate, akin to allowing group-specific intercepts:

$$h(t) = h_{0_q}(t)\exp(\beta' X) \tag{2}$$

$h_{0_q}(t)$ permits a unique baseline hazard for each group or stratum q, distinguishing Equation 2 from Equation 1. When considering whether to stratify the baseline hazard, practitioners' substantive knowledge and expectations serve as the starting point. Broadly speaking, stratifying may be appropriate for two reasons (Metzger and Jones, 2016): either (1) the baseline hazard is likely to significantly differ across groups, or (2) to address violations of the PH assumption.

Regarding the first possibility, baseline hazards may differ across groups for a few reasons. First, there may be substantive reasons to expect groups experience an event at different underlying rates. Myriad rationales exist as to why researchers may think this to be the case, but our running example illustrates one of them: D&W note that the GATT's Article 26 creates a less onerous accession process for former colonies, which may affect the rate at which they apply to join the GATT/WTO. Accordingly, D&W argue there is reason to suspect countries' GATT/WTO application rates may differ for former colonies ($q = 1$, say) vs. other countries ($q = 2$). Fredriksson et al. (2007) have a similar motivation in their study of countries' ratification of the Kyoto Protocol. They stratify by whether a country falls under the Protocol's Annex 1 because Annex 1 vs. non-Annex 1 countries "have widely different responsibilities under the Protocol, which may affect the likelihood of ratification" (Fredriksson et al., 2007, 232).

Second, baseline hazards might differ across groups if subjects are at risk of experiencing more than one transition/event type. Here, the "different groups" amount to potential transition events (e.g., competing risks models). For instance, consider a hypothetical study of applying to *any* trade

[1] For more, see Keele (2010), but see also Metzger (2023).

https://doi.org/10.1017/psrm.2025.12 Published online by Cambridge University Press

organization, perhaps to investigate whether certain geopolitical factors have the same effect on applying regardless of whether the organization is global or regional in scope. In such a scenario, we might consider stratifying the baseline hazard by the organization's scope, to reflect that the underlying rate at which countries apply to join global institutions may differ from the underlying rate at which they apply to join regional institutions. Repeated events models also employ a similar logic for stratifying based on number of events experienced so far, as we discussed earlier.

Our second major reason for stratifying pertains to violating other model assumptions. Specifically, researchers might stratify to address PH assumption violations (Singer and Willett, 2003, sec. 15.2). We discuss this possibility more in the next subsection, but briefly, it may be that two baseline hazards are statistically indistinguishable from one another on average, but if a covariate violates the PH assumption, the two hazards may nevertheless vary significantly in non-proportionate ways over time. Stratifying on that covariate would address the issue. Here, groups would amount to observations with the same value for a PH-violating covariate.

Regardless of the motivation for stratifying, researchers can empirically assess whether doing so is appropriate (Metzger and Jones, 2016, 469). Researchers should estimate an ancillary model that does not stratify the baseline hazard (what we term a "collapsed baseline hazard"), but instead, includes a series of dichotomous variable(s) reflecting the possible groups implied by the stratification variable.

For our running GATT/WTO example with its two strata, we would estimate a Cox model with a single baseline hazard, but with one additional dichotomous covariate indicating whether a country is subject to Article 26.[2] The ARTICLE26 coefficient captures whether the GATT/WTO application rate differs across the two groups. A statistically significant coefficient would be sufficient evidence to stratify.

An insignificant ARTICLE26 coefficient would require us to take an additional step. We would then need to check whether ARTICLE26 violates the PH assumption. Together, these two steps can guide researchers' decisions about where stratification may be appropriate—either because (1) the baseline hazard significantly differs across groups or transitions, or (2) a particular variable violates the PH assumption. Stratification is likely unnecessary if researchers find (a) no significant differences across groups (via ARTICLE26's coefficient) *and* (b) ARTICLE26 does not violate the PH assumption.

1.1.2. Why might order matter?

The number of observed events in a riskset affects the Cox model's statistical power (Hsieh and Philip, 2000; Rosner, 2015, 835–37; Schoenfeld, 1983). Each stratum–(observed failure time) pairing constitutes a riskset, meaning that additional strata correspond to the same number of observations being divided across additional risksets, with fewer total observations in each riskset compared to an unstratified model. Fewer observed events per stratum shrinks the values of the model's expected information matrix (Therneau and Grambsch, 2000, sec. 3.6.1), the inverse of which yields the model's variance–covariance matrix. An information matrix with smaller values thus yields a variance–covariance matrix with larger values—i.e., larger standard errors. The end result is a decreased ability to detect significant non-zero effects, relative to a situation where strata have more observed events.

Collapsing the baseline hazard amounts to pooling together two or more risksets. In a situation with truly collapsible baseline hazards (e.g., if the underlying rate at which Article 26-eligible countries apply to join the GATT/WTO is no different from Article 26-ineligible countries), any assumption checks run after collapsing the baseline hazards would be performed on risksets with more observed events, relative to performing the same tests, pre-collapse. As a result, any tests performed after collapsing the baseline hazards would be better powered, potentially affecting the conclusions researchers reach after checking for PH violations or collapsible covariate effects.

[2] The model specification would be similar for our applying-to-all trade-organizations example. We would need a model with a single baseline hazard, but also with a dichotomous variable indicating whether a particular observation pertains to applying to a global or regional organization.

https://doi.org/10.1017/psrm.2025.12 Published online by Cambridge University Press

1.2. The PH assumption

1.2.1. What is it?

The PH assumption states that the effect of each covariate in the model is proportionate across time. In other words, if a one-unit increase in x's value at $t = 1$ reduces an event's risk of occurring by 20%, that same one-unit increase in x should reduce the event's risk of occurring by 20% at all time points. Covariates violating the PH assumption do not have this proportionate effect across time. There are several ways non-proportionality can occur, but in political science, we usually focus on non-proportionality stemming from covariate effects being conditional on t. For instance, in some models, D&W find evidence that a country's UN voting similarity to the US accelerates its application to join the GATT/WTO, but that this effect declines somewhat as more time passes since the country first becomes eligible for membership (Davis and Wilf, 2017, supplemental appendix, p. 3).

1.2.2. Why might order matter?

The Schoenfeld residual-based test for PH violations is the most common in political science, following the advice of Box-Steffensmeier and Zorn (2001). Schoenfeld residual-based tests can and do diagnose PH assumption violations, but their results can be adversely affected when other issues are present in the data, such as model misspecification (Keele, 2010; Metzger, 2023; Therneau and Grambsch, 2000, sec. 6.6). Making an incorrect conclusion about stratification or whether to estimate stratum-specific covariate effects are examples of misspecification.

Moreover, estimates of x's effect may be biased if x violates the PH assumption and we do not correct for the violation, making any tests we perform using these biased estimates potentially incorrect. If we continue to conceive of PH violations as situations in which x's effect is conditional on t's value, from broader work on interaction terms, we know that failing to account for x's conditional effect produces a $\widehat{\beta}_x$ estimate equal to the weighted average of x's conditional effect across the conditioning variable's values (Brambor et al., 2006, 73).[3] If we check whether a covariate's stratum-specific effects can be collapsed *before* checking for PH, we would be using this weighted conditional effect when doing this check—a potentially dangerous situation for inference, if (a) x violates PH for some transitions, but not all, or (b) the value of x's t-conditional effect differs across transitions (e.g., it is positive for some transitions but negative for others), even if x violates PH for every transition.

1.3. Collapsed versus stratum-specific covariate effects

1.3.1. What are they?

Stratum-specific covariate effects allow the same variable to have a different effect (coefficient estimate) for each stratum in the model (Equation 3). The subscripts on the coefficient vectors indicate that, for each of our covariates, we do not constrain its coefficient to be equal across every stratum $q \in \{1,\ldots,Q\}$.

$$h(t) = h_0(t)\exp(\beta'_{q_1}X + \beta'_{q_2}X + \cdots + \beta'_{q_Q}X) \tag{3}$$

If we extend our running GATT/WTO example to applying to join any trade organization, we would estimate two distinct sets of coefficients, reflecting the possibility that the same covariate may have a different effect on applying to a global versus regional organization. Including stratum-specific effects is standard practice in multistate models and competing risks models, which are a special case of multistate models. Importantly, though, any duration model with strata can employ this specification strategy. For instance, with D&W's stratified Cox model, we might consider—as we do later—if the effects of political and economic factors on time to apply to the GATT/WTO vary depending on whether a country is subject to Article 26.

[3] In Cox models, this weighted average also depends on t's magnitude and the dataset's distribution of right-censored values across t (Horiguchi et al., 2019).

https://doi.org/10.1017/psrm.2025.12 Published online by Cambridge University Press

Researchers can assess whether stratum-specific effects are appropriate by including them in the model specification, estimating the model, and then using Wald and/or likelihood-ratio tests to see whether a covariate's effect is distinguishable across two or more strata (Metzger and Jones, 2016). If the stratum-specific effects do not statistically differ from one another across some strata for a covariate, we can estimate fewer separate effects for that covariate, to the point that we can estimate a single collapsed effect if the covariate's effects do not differ across any of the strata.

Notably, estimating stratum-specific covariate effects involves more estimated parameters than collapsed covariate effects. If no stratum-specific effects exist, including them will be inefficient because x's effect could be modeled with fewer estimated parameters. However, prematurely collapsing a covariate's effect when stratum-specific effects exist can produce biased estimates.

1.3.2. Why might order matter?

We previously noted why ordering might matter for collapsed coefficient effects when discussing the other assumptions, such as statistical power (Section 1.1.2) or masking (or inducing the appearance of) PH violations (1.2.2).

We now turn to assessing whether our concerns about test sequencing bear out in practice using simulations.

2. Simulation setup

2.1. Motivations

We design our simulations to investigate the order in which to test these assumptions under the best of circumstances. We deliberately use a large n (=2000) to ensure any asymptotic test properties are active and that our tests will be sufficiently powered.[4] As we discuss in the next subsection, our simulated process only has two possible transitions (synonymously, event types or strata), and each transition's hazard is affected by only two uncorrelated covariates—the simplest possible structure we could use to address our questions of interest. This structure is identical to our running D&W example. D&W's belief that Article 26-eligible countries will likely have different application rates than other countries is an implicit hypothesis about two groups having different baseline hazards, yielding two distinct transitions/strata: one for Article 26-eligible countries and one for Article 26-ineligible countries.

We approach our investigation from the perspective of a hypothetical practitioner who has substantively motivated hypotheses about both (1) two or more groups having different baseline hazards and (2) two or more groups having different covariate effects for a common list of covariates, and is worried about whether the order in which s/he assesses these hypotheses could affect his/her conclusion.[5,6] D&W's analysis nearly fits this description. Extending D&W's suspicions about Article 26 countries to their main covariates of interest—applicant countries' geopolitical similarity with GATT/WTO members—would also give us implicit hypotheses about different covariate effects. Specifically, geopolitical similarity with GATT/WTO members may have a different effect on Article 26-eligible countries' application rates compared to Article 26-ineligible countries. However, our simulation results also speak to practitioners who may have some hypotheses about either different baseline hazards *or* different covariate effects, but not both.

[4] Power considerations would impact the performance of these tests, in practice. When we rerun our simulations with $n = 100$, performance is poor across the board. As a consequence, we opt to hold power concerns constant in our investigation, allowing us to assess the optimal test sequence in best-case scenarios before introducing additional factors that may further impact such decisions.

[5] We also run a condensed set of simulations for Orderings A and B in which we record the final covariate estimates implied by each ordering, to spot check for estimate bias. By and large, the coefficient estimates are unbiased in aggregate, though the amount of bias is not necessarily negligible in value; see Appendix E for details.

[6] We are agnostic as to whether the estimates in our final specifications reflect causal effects. Speaking to causality involves broader features of the research design beyond model specification.

https://doi.org/10.1017/psrm.2025.12 Published online by Cambridge University Press

2.2. Technical details

We structure our simulations such that all subjects are at risk of experiencing only one of the two transitions.[7] Once a subject experiences its assigned transition, it is no longer at risk of experiencing any transitions. Aside from its similarity to D&W's structure, we opted for this general structure because, as we discuss later, it gives us complete control over (a) the number of subjects at risk of each transition (i.e., riskset size), which in turn impacts (b) the number of subjects that experience their riskset's transition.

To create this structure, we use survsim (Crowther and Lambert, 2012; Crowther, 2022) to generate data from a process composed of two mutually exclusive transitions. Each transition q's data-generating process (DGP) is a Weibull hazard: $h_q(t) = h_{0_q}(t)\exp(\beta_q' X + \phi_q' g(t) X)$, where the Weibull's $h_{0_q}(t) = \lambda_q p_q t^{(p_q - 1)}$, λ_q is transition q's scale parameter, and p_q is transition q's shape parameter. $g(t)$ represents some function of t on which the covariates' effects are conditioned, and $\beta_q' X$ and $\phi_q' g(t) X$ represent the covariates and their main effects (β_q) and their (potentially) time-varying effects (TVEs, ϕ_q) on transition q's hazard. All our PH violations, when present, use $g(t) = \ln(t)$. We include two uncorrelated covariates for both transitions' hazards, x_1 and x_2, both distributed normal with mean 0 and standard deviation 0.25.

We use Cox models as our duration model of choice. Other simulation studies involving the Cox have shown the number of subjects experiencing a transition matters, both for estimating the model (e.g., presence of many tied survival times [Hertz-Picciotto and Rockhill, 1997]) and for being able to detect effects (Hsieh and Philip, 2000; Rosner, 2015, 835–37; Schoenfeld, 1983). Our process' structure, with subjects at risk of one transition only, allows us to adjust this number directly; we would be unable to do so with other structures, such as a competing-risks or repeated-events setup, because characteristics of the *other* transitions' hazards ($\neg q$) also affect whether a subject experiences q. All and all, our structure's simplicity gives us control over the number of subjects experiencing each transition, but simultaneously, allows us to introduce transition-specific baseline hazards and/or coefficient effects while bracketing the other, more complex dynamics that characterize competing-risks or repeated-event structures. This is another way in which our setup constitutes a best-case scenario: the only factors affecting whether subjects experience q are related only to q's hazard and any amount of right censoring we exogenously impose.

We randomly assign 2000 subjects to one of our two transitions, irrespective of the subjects' covariate values. For each of our scenarios, we divide the 2000 subjects between the two transitions in five ways: {10%, 90%}, {30%, 70%}, {50%, 50%}, {70%, 30%}, {90%, 10%}.[8] We impose no right censoring, meaning the number of subjects capable of experiencing transition q will equal the number of observed transition-q events. Again, these choices represent a best-case scenario: any asymptotic test properties will be active with $n = 2000$, and our data are informative as possible because we fully observe all subjects' failure times.

Four basic patterns characterize our DGPs, involving (a) whether the baseline hazard, $h_0(t)$, varies by transition, (b) whether the covariates have transition-specific effects, (c) whether the covariates violate PH, and (d) if they violate PH, whether the violation differs by transition. When we induce violations of our various assumptions, we generally induce relatively large violations, to further bracket any power concerns. We introduce transition-specific baseline hazards by varying λ_2's value in majority of the scenarios. We opted to manipulate λ_2 to minimize any impact on our ability to potentially detect PH violations, because we always have PH-violating covariates for transition 1, but this is not always the case for transition 2. Table 1 lists our four major patterns regarding x_1's effects (first two columns), along with the parameters whose values change across scenarios, while

[7] See Appendix B's Figure 4 for a stage diagram.

[8] The 10/90 and 90/10, and 30/70 and 70/30, transition distributions will not produce identical results. The DGPs for transitions 1 and 2 are different, meaning that 30% of subjects at risk of transition 1, e.g., will not produce equivalent results as 30% of subjects being at risk of transition 2.

https://doi.org/10.1017/psrm.2025.12 Published online by Cambridge University Press

Table 1. Four major patterns: varying parameter values

		Main	*TVE*	*Scale*
Description (re: x_1 effects)	Scen.	x_{1_tr1}	x_{1_tr2}	λ_2
Same main, tr1 TVE only	1	0.65	0	0.005
	5	0.65	0	0.02
	9	0.65	0	0.05
	13	0.65	0	0.1
Same main, same TVE	2	0.65	−0.45	0.005
	6	0.65	−0.45	0.02
	10	0.65	−0.45	0.05
	14	0.65	−0.45	0.1
Same main, diff. TVE	3	0.65	0.45	0.005
	7	0.65	0.45	0.02
	11	0.65	0.45	0.05
	15	0.65	0.45	0.1
Diff. main, diff. TVE	4	−0.65	0.45	0.005
	8	−0.65	0.45	0.02
	12	−0.65	0.45	0.05
	16	−0.65	0.45	0.1

Note: TVE = time-varying effect. TVE's true time transformation = ln(t). Darkest gray shading: the transitions' baseline hazards are identical.

Appendix B's Table 4 contains the parameters whose values remain constant across scenarios.[9] Coupled with our five different subjects-per-transition distributions, we estimate 100 scenarios in total and run 1000 simulation draws for each.

Our four basic patterns allow us to assess how each ordering performs in the presence or absence of particular violations. Appendix B reports our simulations' specific parameter values, but the patterns can be summarized thus:

1. Collapsible $h_0(t)$: Scs. 9–12 only
2. PH violations
   a. x_1
      i. Transition 1: all scenarios
      ii. Transition 2: all scenarios except those in which dividing the scenario ID by 4 yields a remainder of 1.
   b. x_2: never violates for either transition
3. Collapsible covariate effects
   a. x_1
      i. Main effect: all scenarios except those that are multiples of four
      ii. TVE: all scenarios that are multiples of two but *not* multiples of four
   b. x_2
      i. Main effect: never collapsible
      ii. TVE: *N/A (never violates PH)*

2.3. *Draw-by-draw procedure*

For each draw, we start by estimating a Cox model in the typical multistate form, meaning that we permit transition-specific baseline hazards and transition-specific effects for both covariates, for the reasons we discussed in Section 1. We then begin checking Section 1's three featured assumptions.

There are six possible ways in which we could order the three assumption checks (Table 2). We run all six possibilities on every draw. We suspect orderings that check for PH violations before checking for collapsed covariate effects (Table 2's lightly shaded columns) will tend to arrive at the correct

[9] Also see Appendix B for a single table containing both tables' values.

https://doi.org/10.1017/psrm.2025.12 Published online by Cambridge University Press

Table 2. Possible test orderings

A	B	C	D	E	F
$h_0(t)$ $\hat{\beta}$ PH test	$h_0(t)$ PH test $\hat{\beta}$	$\hat{\beta}$ $h_0(t)$ PH test	$\hat{\beta}$ PH test $h_0(t)$	PH test $h_0(t)$ $\hat{\beta}$	PH test $\hat{\beta}$ $h_0(t)$

Key: $h_0(t)$ = check whether baseline hazards are collapsible; $\hat{\beta}$ = check whether a coefficient's effects are collapsible; PH test = check whether a covariate violates the PH assumption. Shaded orderings: checks for PH violations before $\hat{\beta}$ collapse.

final specification more frequently than other orderings, across a greater number of scenarios, for the reasons we discussed in Section 1.2.2. Accordingly, we anticipate Orderings B, E, and F will arrive at the correct final model specification more often than Orderings A, C, and D because B/E/F do not collapse any estimated effects until *after* we check and correct for any PH violations.

We report the descriptive results for all six orderings in Appendix F. The results suggest focusing on only Orderings A and B in the main text is sensible, as we discuss in that appendix.

Because we focus on A and B, we can be more specific about when we check which assumption. First, we check whether we can collapse the baseline hazard. We update our specification based on this check's output. Using the updated specification, we then vary the order in which we check (1) whether we can collapse a given covariate's effect across the two transitions and (2) whether any covariates violate PH for any transition.[10] We use $\ln(t)$ as the time transform for our Schoenfeld PH tests, matching the true DGP's TVE time transform. When we find PH violations, we correct for the violation by including an interaction between the offending covariate and $\ln(t)$.

- Ordering A ($h_0(t)$, $\hat{\beta}$, PH): We check whether we can collapse the transition-specific covariate effects, reestimate the model again using the specification implied by the Wald output, then check for any PH violations. We implement any corrections for PH violations and treat this as the ordering's final specification.
- Ordering B ($h_0(t)$, PH, $\hat{\beta}$): We check for PH violations, reestimate the model with any necessary PH corrections, then check to see whether we can collapse covariate effects. We check both the covariates' main effects and TVEs, if a given covariate is found to violate PH for both transitions. Thus, we assess not only whether the covariate's main effects are equivalent across transitions, but also whether its TVEs are equivalent across transitions. We collapse any effects based on the Wald output and treat this as the ordering's final specification. As we discussed earlier, we anticipate this ordering will uncover the correct final specification more frequently than Ordering A because it does not collapse any estimated covariate effects until after we check and correct for any PH violations.

We record the final specification for both orderings. Assessing which ordering does better, then, is simply a matter of counting how many times its final specification matches the true specification for a given transition. We classify a specification as correct for transition q if it matches q's true DGP for main effects and TVEs in all respects (see Appendix D).

3. Simulation results

Figure 1 contains the simulation results for Scenarios 1–16 from our {50%, 50%} distribution. We report our results as a series of horizontal stacked bar graphs, displayed as four panels, arranged from the smallest λ_2 value at the figure's upper left to the largest λ_2 value at its lower right. For each

[10] There have also been shifts in how some packages implement the PH test's calculations. Current evidence suggests the newer implementation (currently, R only) performs better (Metzger, 2023), but importantly, only in situations where the covariates are uncorrelated (Metzger, 2024). All our reported results use the traditional implementation.

https://doi.org/10.1017/psrm.2025.12 Published online by Cambridge University Press

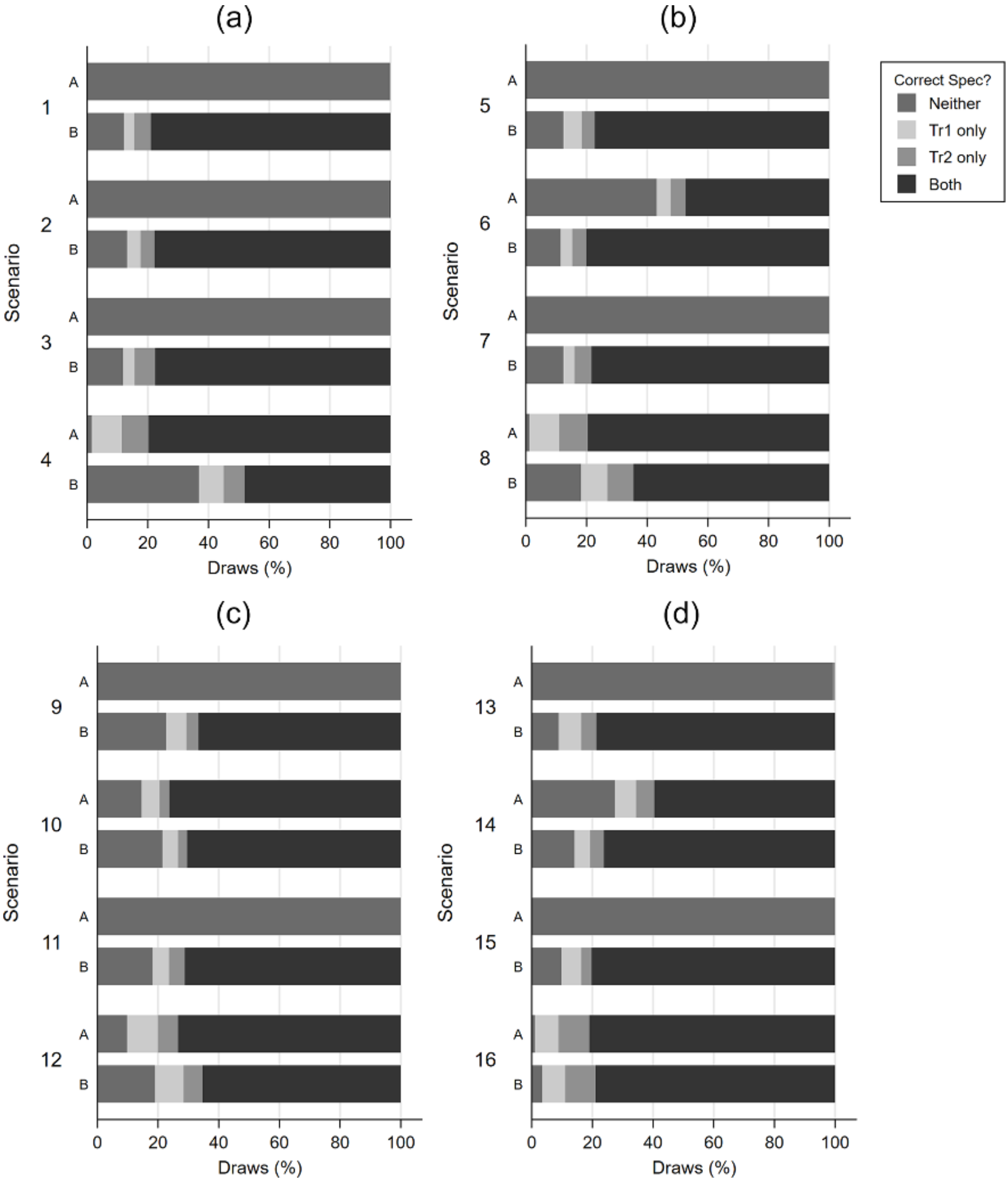

Figure 1. Simulation results (a) $\lambda_2 = 0.005$, (b) $\lambda_2 = 0.02$, (c) $\lambda_2 = 0.05$ and (d) $\lambda_2 = 0.1$
$n = 2000$, subjects distributed 50%/50% across two transitions, 1000 simulations. Test Ordering: A: $h_0(t)$ collapse, $\hat{\beta}$ Wald test, PH test. B: $h_0(t)$ collapse, PH test, $\hat{\beta}$ Wald test.

scenario, we break down an ordering's performance into four mutually exclusive possibilities, based on the final specification it arrives at for each transition: (1) the ordering's final specification gets neither transition's true DGP correct (leftmost segment of bar); (2) and (3) the ordering gets one of the transition's DGPs correct, but not the other (middle two segments, for transition 1 and transition 2, respectively); and (4) the ordering gets both transitions' DGPs correct (rightmost segment, darkest gray shade). If both orderings always recover the correct final model specification, we should see bars filled with the darkest of the four gray shades.

https://doi.org/10.1017/psrm.2025.12 Published online by Cambridge University Press

Figure 1 makes clear that neither ordering performs perfectly. However, they are not equally imperfect. Ordering matters in a non-trivial way—an important finding. In the first three scenarios in every panel, save one (Scenarios 1–3, 5–7, 9, 11, 13–15), Ordering B outperforms Ordering A by a large margin, evident in the way Ordering B's darkest gray segment (both transition specifications correct) is appreciably larger than Ordering A's. These scenarios correspond to x_1 having the same main effect across both transitions, speaking to whether we can correctly detect collapsible covariate effects across transitions using the Wald test. The slimmest margin by which Ordering B outperforms Ordering A in Figure 1 is 16.6 percentage points (Sc. 14), while the largest margin is 80.3 percentage points (Sc. 15).

We find Ordering A's generally poor performance in our scenarios stems from collapsing covariate effects first. Doing so hinders the PH test's ability to detect PH violations, as we suspected it would. It is also no coincidence that Ordering A performs the worst in the first and third scenarios in each panel. In these scenarios, x_1's main effect can be collapsed, but its TVE cannot, as it differs by transition. x_1's transition-specific TVEs produce the appearance of non-equal x_1 main effects (before correcting for the PH violations), leading A to conclude erroneously that x_1's main effect cannot be collapsed, immediately making A's specification for both transition 1 *and* transition 2 incorrect.

The orderings' relative performance becomes murkier once x_1's main effect is no longer collapsible in truth (final scenario in all panels). Ordering A now outperforms Ordering B in each panel, albeit by an increasingly small margin as λ_2 increases. In Figure 1's top-left panel, Ordering A arrives at the correct specification 79.8% of the time in Scenario 4—31.7 percentage points higher than Ordering B. In looking into Ordering B's performance, B tends to arrive at one of two specifications that differ in their conclusions about x_1's main effect when λ_2 is small (Sc. 4). In one specification (call this ♣), x_1's main effect is not collapsed, correctly (48.1% of draws). In the other (♠), x_1's main effect is incorrectly collapsed (31.7% of draws). Both ♣ and ♠ reach the correct conclusion for x_2's main effect (not collapsed) and x_1's TVE (both transitions violate and not collapsed). However, as λ_2 increases in value, Ordering B arrives at the correct specification with greater frequency (in Sc. 16: ♣ = 79.0%, vs. ♠'s 2.0%), to the point where Ordering A outperforms B only by 2 percentage points in Sc. 16 (bottom-right panel).

From examining each scenario's individual draws, Ordering B's shaky performance when x_1's effect is no longer collapsible mainly derives from standard error sizes. Modeling the PH effects first tends to enlarge the standard errors corresponding to the PH-violating covariate's main effect, as is the case generally for any interaction term's constituent parts. This matters when λ_2 is small, because the standard errors tend to be enlarged as is, reducing the Wald test's subsequent ability to distinguish between x_1's main effects across transitions. By contrast, Ordering A's biggest liability in the panels' other three scenarios becomes its biggest asset here—by not checking for PH violations first, there are no PH violation corrections, resulting in smaller standard errors for x_1's main effects, which makes it easier for Ordering A's Wald test to correctly detect that x_1's main effect is statistically different across the two transitions and should not be collapsed.

These broad patterns hold across the other four subjects-per-transition distributions,[11] in terms of A vs. B's relative performance in each scenario.[12] This means the performance of Ordering A vs. B is not so disproportionally affected by the number of subjects assigned to each transition that the orderings' relative performance changes. In absolute terms, both orderings tend to have a harder time uncovering the correct specification for both transitions when few subjects are assigned to any transition, compared to when subjects are equally distributed across transitions.[13] These same

[11] See supplemental viewing app to easily view all our results (https://tinyurl.com/muazzaem).

[12] The only exception is Sc. 14, 90%/10%, where A and B perform nearly identically (correct specification for both transitions: 28.7% [A] vs. 28.2% [B]). B otherwise outperforms A by a large margin in the other four subject–transition distributions.

[13] These absolute performance differences are not mirrors on either side of the 50%/50% distribution, for reasons discussed in fn. 8.

https://doi.org/10.1017/psrm.2025.12 Published online by Cambridge University Press

patterns also broadly hold if we vary p_1's value (Scenarios 1009–1012), as another way to make the transitions' baseline hazards unequal (Appendix C, for {50%/50%}).

Our results are scoped to the scenarios we examine, as the results of any (Monte Carlo) experiment are, but they serve as an important benchmark. They show that ordering can matter—a notable finding in its own right—and suggest which orderings tend to work better under optimal conditions, providing practitioners with important guidance. That our results mirror our ex-ante suspicions makes us cautiously optimistic about their generalizability to other, more complex DGPs.

4. Application

To demonstrate our simulation results' substantive implications, we reexamine D&W's analysis of the timing of countries' GATT/WTO membership applications. Prevailing understandings of applying to free trade organizations emphasize economic determinants, such as trade preferences and relations with existing members. In contrast, D&W contend that the decision to apply is largely driven by geopolitical considerations, such as democracy, alliances with current members, and preference similarity with the US as reflected in UN voting patterns. Given the dependent variable, D&W use a Cox model to test their hypotheses.

Importantly, the authors note Article 26 of the GATT creates a unique, less onerous GATT accession process for former colonies of member countries, which may produce a distinct rate at which former colonies apply. As a result, D&W stratify their Cox model based on whether a country is eligible for GATT entry under Article 26, permitting each group to have its own baseline hazard. Following convention, the authors estimate a single coefficient for each covariate in the model, making the implicit assumption that these variables' effects will be the same regardless of a country's Article 26 eligibility. They then test for PH assumption violations using these collapsed coefficients and make corrections where appropriate.

We can replicate D&W's Table 1, Model 3 without issue (Table 3, first model). Consistent with the authors' original results, political factors, such as whether a country is a democracy, its UN voting similarity with the US, and the number of allies it has in the GATT/WTO accelerate the rate at which countries apply. Economic factors such as trade openness fail to achieve statistical significance.

To examine the impact of the order in which these tests are performed, we begin by re-estimating D&W's model, but allow the effect of each covariate to vary across the model's two strata, analogous to our simulation setup. We first evaluate whether a stratified model is appropriate, as Orderings A and B do, by including a dichotomous indicator of whether a country is Article 26 eligible as a predictor in the model. We find this variable is *not* a significant predictor of GATT/WTO application timing, nor does it violate the PH assumption. Consequently, we conclude it is not necessary to stratify the baseline hazard by Article 26 eligibility.

Next, we test whether we can collapse these coefficient effects and for PH assumption violations: once using Ordering A ($h_0(t)$, $\hat{\beta}$, PH), and then again using Ordering B ($h_0(t)$, PH, $\hat{\beta}$). We report the final specifications implied by the orderings in Table 3's second and third major columns, respectively.

Ordering A's final specification has several important differences from D&W's. First, our Wald tests indicate most of the model's covariates have a similar effect on GATT/WTO application rates regardless of a country's Article 26 eligibility, aligning with D&W's original specification. However, we find three variables' effects differ based on Article 26 eligibility: UN voting similarity with the US; the number of allies already in the GATT/WTO; and the number of GATT/WTO members (%WorldAreMembers). The first two variables correspond to D&W's general hypothesis about geopolitical factors. For both variables, Ordering A's results indicate they *only* increase the rate at which former colonies apply to join the organization. For all other countries, these political considerations no longer have a significant effect. This suggests an important scope condition to D&W's findings:

https://doi.org/10.1017/psrm.2025.12 Published online by Cambridge University Press

Table 3. Davis and Wilf: replication results

	D&W's T1, M3 (Replication)	Ordering A Specification		Ordering B Specification	
	All	∼Art. 26	Art. 26	∼Art. 26	Art. 26
Polity	0.055***	0.078***	0.078***	0.079***	0.079***
	(0.016)	(0.018)	(0.018)	(0.018)	(0.018)
UN voting similarity	0.012**	0.005	0.020***	0.003	0.027***
	(0.005)	(0.004)	(0.005)	(0.004)	(0.006)
Ally member count	0.060**	0.019	0.173***	0.021	0.172***
	(0.022)	(0.029)	(0.033)	(0.030)	(0.035)
ln(Openness)	0.198	0.313*	0.313*	0.385**	0.385**
	(0.127)	(0.124)	(0.124)	(0.136)	(0.136)
GATT/WTO trade %	−0.073	−0.242	−0.242	0.077	0.077
	(0.670)	(0.636)	(0.636)	(0.641)	(0.641)
ln(GDP)	0.365***	0.424***	0.424***	0.458***	0.458***
	(0.080)	(0.082)	(0.082)	(0.083)	(0.083)
ln(GDPPC)	−0.212^{+}	−0.238*	−0.238*	−0.265*	−0.265*
	(0.125)	(0.114)	(0.114)	(0.112)	(0.112)
Cold War	−1.577***	−1.940***	−1.940***	−2.212***	−2.212***
	(0.444)	(0.384)	(0.384)	(0.550)	(0.550)
Cold War * ln(t)				0.334	
				(0.244)	
Democratizing country?	−0.374	−0.524	−0.524	−0.440	−0.440
	(0.377)	(0.422)	(0.422)	(0.413)	(0.413)
Former colony?	0.208	0.029	0.029	0.178	0.178
	(0.246)	(0.226)	(0.226)	(0.231)	(0.231)
PTA trade %	1.746*	1.649*	1.649*	2.103**	2.103**
	(0.746)	(0.779)	(0.779)	(0.768)	(0.768)
Percent world members	0.919	−2.843	−1.085	−3.703*	−3.703*
	(2.074)	(1.729)	(1.634)	(1.664)	(1.664)
Percent world members * ln(t)					1.189***
					(0.294)
Trade round?	−1.429**	−1.431**	−1.431**	−0.477*	−0.477*
	(0.483)	(0.457)	(0.457)	(0.236)	(0.236)
Trade round? * ln(t)	0.595**	0.545**	0.545**		
	(0.229)	(0.203)	(0.203)		
n	1750	1750		1750	
n_{fail}	120	78	42	78	42
$h_0(t)$: Stratified on Art. 26?	Yes	No		No	

***$p<0.001$.
**$p<0.01$.
*$p<0.05$.
$^{+}p<0.1$, two-tailed values. Standard errors clustered on country in parenthesis.
For Ordering A's and B's specifications, the same coefficient estimate and standard error in a row = collapsed effect, different coefficient estimates = stratum-specific effects (first estm. = Article 26-ineligible countries; second estm. = Article 26-eligible countries).

geopolitical considerations matter for former colonies, but less so for other countries.[14] Second, in Ordering A's final specification, trade openness is now positive and statistically significant, in contrast to D&W's original non-finding for all economic factors.

Ordering B's final specification supports these same broad patterns. Consistent with Ordering A, this specification suggests trade openness is a significant predictor of when a country applies to the GATT/WTO, with countries more open to trade applying more quickly. Moreover, this specification again finds the number of allies a former colony has in the GATT/WTO accelerates time-to-applying, as does UN voting similarity with the US, but these two covariates have no statistically significant effect for other countries.

[14] D&W's third and final geopolitical variable is regime type. It remains positive and significant for *all* potential joiners, in line with Davis and Wilf's original analysis.

https://doi.org/10.1017/psrm.2025.12 Published online by Cambridge University Press

However, Ordering B's model also offers several additional implications. As one example, both Ordering A and B suggest countries applied to the GATT/WTO at a lower rate during the Cold War. But, Ordering B also classifies this variable as a PH violator for Article 26-ineligible countries, implying application rates vary across time for non-colonies during the Cold War, but no such TVE exists for colonies during the Cold War. As an additional example, Ordering A detects Article 26-specific effects for %WorldAreMbers. By contrast, Ordering B finds this covariate has no Article 26-specific main effects, but finds the covariate violates PH for Article 26-eligible countries, suggesting Ordering A not collapsing %WorldAreMbers's effects may stem from missing this PH violation. For D&W's application, the two orderings happen to reach broadly similar conclusions. Nevertheless, we would place greater weight on Ordering B's final specification, on the basis of Ordering B outperforming Ordering A in the bulk of our simulations.

The scope condition we identify for D&W's geopolitical covariates is additionally interesting in light of our findings about the baseline hazard being collapsible. It suggests that, despite the differences in accession criteria for Article 26-eligible countries, they still fundamentally apply to join the GATT/WTO at the same underlying rate as Article 26-ineligible countries, *once we allow* various geopolitical and economic factors to have stratum-specific effects. Instead, geopolitical factors are the main reason Article 26-eligible countries appear to apply at a different rate than Article 26-ineligible countries. These Article 26-specific covariate effects may be why D&W found evidence supportive of stratifying for their original model[15]—and, indeed, if we rerun D&W's original model with its collapsed covariate effects and assess whether stratifying on Article 26 is necessary using the procedure we describe in Section 1.1.1, the evidence supports stratifying.

5. Conclusion

In this article, we use Monte Carlo simulations to assess how frequently we reach the correct model specification for a Cox duration model using different specification test orderings. We focus on specification decisions pertaining to stratified hazards, the PH violation, and collapsed covariate effects. Our simulations use the same starting specification as competing risks and multistate models. They also constitute best-case scenarios, to provide a sense of how frequently we reach the correct model specification in optimal conditions, as we suspect this frequency will be far lower in the less-optimal conditions that characterize real data.

Broadly speaking, our results indicate the possible orderings cluster into two groups of equal size, when it comes to performance similarities. Among the scenarios we examine, the best-performing orderings are those that, first, test and correct for PH assumption violations, then check whether any covariate effects are collapsible. Ordering B is an example: it checks for collapsible baseline hazards first, then PH violations, then collapsible covariate effects. These orderings tend to reach the correct specification most frequently—almost twice as frequently, in aggregate, compared to the cluster of poorer-performing orderings (e.g., Ordering A [collapsible baseline hazards, collapsible covariate effects, PH violations]).

We also break down Ordering A vs. B's performance in more detail, as these are the two orderings we feel practitioners will gravitate toward using because they are easy to implement. Here, we find Ordering B typically outperforms Ordering A, as the aggregate results suggest, with a few exceptions. Using our simulation results as a springboard, our reexamination of Davis and Wilf (2017) reveals their key covariates do matter, but only for a subset of countries, showcasing the importance of both adopting a permissive starting model specification and being mindful of the order in which we subsequently assess these more permissive specification decisions.

As duration models become more widely adopted in political science, researchers are using increasingly complex model variants. This complexity has allowed researchers to more closely tailor

[15] D&W used a log-rank test to check whether stratifying was necessary.

https://doi.org/10.1017/psrm.2025.12 Published online by Cambridge University Press

their statistical models to their theoretical concerns by, for example, allowing covariate effects to vary depending on the context in which they are observed, or allowing for covariate effects to change over time. However, it has also increased the number of specification tests researchers should perform. Until now, there has been relatively little guidance for researchers about the order in which to perform these tests. Our article provides some of the first systematic guidance in this regard, which should be of particular interest to researchers with substantive hypotheses implying different baseline hazards and/or different covariate effects across groups. We anticipate the importance of performing diagnostic tests in the optimal order to increase as the use of more complex model variants grows.

Supplementary material. To view supplementary material for this article, https://doi.org/10.1017/psrm.2025.12. To obtain replication material for this article, https://doi.org/10.7910/DVN/XZHXVF.

Acknowledgements. The authors' names appear in alphabetical order. This article was presented at the 2021 annual meeting of the Midwest Political Science Association and American Political Science Association. We thank Robert Franzese and Harvey Palmer for feedback, as well as the IR/Comparative reading group at the University of Mississippi. We also thank David Darmofal for feedback on an earlier, related version of this project. We bear sole responsibility for errors and shortcomings.

Conflicts of interest. The authors declare none

Funding. None

Data availability. All analyses are performed using Stata 17.0/R 4.4.1. This article's replication code is available via Dataverse at https://doi.org/10.7910/DVN/XZHXVF

References

Blair CW, Chenoweth E, Horowitz MC, Perkoski E and Potter PBK (2022) Honor among thieves: Understanding rhetorical and material cooperation among violent nonstate actors. *International Organization* **76**(1), 164–203.

Box-Steffensmeier JM and Zorn CJW (2001) Duration models and proportional hazards in political science. *American Journal of Political Science* **45**(4), 972–988.

Brambor T, Roberts Clark W and Golder M (2006) Understanding interaction models: Improving empirical analyses. *Political Analysis* **14**(1), 63–82.

Carter J and Lemke D (2022) Birth legacies and state failure. *Journal of Conflict Resolution* **66**(10), 1854–1880.

Conrad CR and Moore WH (2010) What stops the torture? *American Journal of Political Science* **54**(2), 459–476.

Crowther MJ (2022) Simulating time-to-event data from parametric distributions, custom distributions, competing-risks models, and general multistate models. *Stata Journal* **22**(1), 3–24.

Crowther MJ and Lambert PC (2012) Simulating complex survival data. *Stata Journal* **12**(4), 674–687.

Davis CL and Wilf M (2017) Joining the club: Accession to the GATT/WTO. *Journal of Politics* **79**(3), 964–978.

Fredriksson PG, Neumayer E and Ujhelyi G (2007) Kyoto protocol cooperation: Does government corruption facilitate environmental lobbying? *Public Choice* **133**(1/2), 231–251.

Gelman A and Loken E (2014) The statistical crisis in science. *American Scientist* **102**(6), 460–465.

Hertz-Picciotto I and Rockhill B (1997) Validity and efficiency of approximation methods for tied survival times in Cox regression. *Biometrics* **53**(3), 1151–1156.

Hsieh FY and Philip WL (2000) Sample-size calculations for the Cox proportional hazards regression model with nonbinary covariates. *Controlled Clinical Trials* **21**(6), 552–560.

Jones BT and Metzger SK (2018) Evaluating conflict dynamics: A novel empirical approach to stage conceptions. *Journal of Conflict Resolution* **62**(4), 819–847.

Keele L (2008) *Semiparametric Regression for the Social Sciences*. New York: Wiley.

Keele L (2010) Proportionally difficult: Testing for nonproportional hazards in Cox models. *Political Analysis* **18**(2), 189–205.

Maeda K (2010) Two modes of democratic breakdown: A competing risks analysis of democratic durability. *Journal of Politics* **72**(4), 1129–1143.

Metzger SK (2023) Proportionally less difficult?: Reevaluating Keele's 'Proportionally difficult.' *Political Analysis* **31**(1), 156–163.

Metzger SK (2024) Implementation matters: Evaluating the proportional hazard test's performance. *Political Analysis* **32**(2), 240–255.

Metzger SK and Jones BT (2016) Surviving phases: Introducing multistate survival models. *Political Analysis* **24**(4), 457–477.

Miki H, Hassett MJ and Uno H (2019) How do the accrual pattern and follow-up duration affect the hazard ratio estimate when the proportional hazards assumption is violated? *The Oncologist* **24**(7), 867–871.

Rosner B (2015) *Fundamentals of Biostatistics*, 8th ed. Boston: Cengage Learning.

Schoenfeld DA (1983) Sample-size formula for the proportional-hazards regression model. *Biometrics* **39**(2), 499–503.

Singer JD and Willett JB (2003) *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*. Oxford: Oxford University Press.

Therneau TM and Grambsch PM (2000) *Modeling Survival Data: Extending the Cox Model*. New York: Springer.

Wong SH-W and Chun-Man Chan K (2021) Determinants of political purges in autocracies: Evidence from ancient Chinese dynasties. *Journal of Peace Research* **58**(3), 583–598.

Cite this article: Jones BT and Metzger SK (2025) Out of order: specification check sequencing in Cox models. *Political Science Research and Methods*, 1–17. https://doi.org/10.1017/psrm.2025.12