Appendices for "Getting Time Right," by Shawna K. Metzger and Benjamin T. Jones (*Political Analysis*)

Appendix A:
Cox/cloglog Equivalence Derivation

In this appendix, we reproduce the derivation demonstrating that a continuous-time Cox semi-parametric duration model is equivalent to a discrete-time cloglog model with time dummies for its temporal dependence correction (also referred to as “grouped duration data”). We include the equivalence derivation for reference purposes; others have shown the same elsewhere (e.g., Beck, Katz, and Tucker 1998, 1284–85; Cameron and Trivedi 2005, 600–601, 602–3). We focus on the general intuition behind the derivation, leaving the underlying mathematics to the previously cited others. We try to use *j* to denote calendar time (e.g., year) for an observation, and continue using *t* for the duration.

The place to begin is by realizing that a continuous Cox model is usually expressed in terms of *h*(*t*), the hazard, and that a cloglog model (discrete time or otherwise) is usually expressed in terms of Pr(y = 1), the probability of an event occurring. When we speak of discrete-time durations, $y_{ij} = 1$ means the duration terminates in time period *j* for subject *i*. The cloglog’s functional form is (Long 1997, 51):

$$\Pr(y = 1) = 1 - \exp[-\exp(\beta' X)] \qquad 1$$

To show equivalence, we will need to first reexpress the continuous-time Cox’s hazard function in terms of a probability. We can then convert the continuous-time Cox expression into discrete time, and see if the resultant expression is equivalent to equation 1.

A. Continuous Cox: Hazard to Probability

We start from *h*(*t*):

$$h(t) = h_0(t) \exp\left(\beta' X_{ij}\right) \qquad 2$$

From *h*(*t*), we can derive the expression for *S*(*t*), the survivor function, because writing an expression for *h*(*t*) necessarily specifies a distribution for *t* (Kalbfleisch and Prentice 2002, 7). If we know *t*’s distribution, any function involving *t*’s distribution—its probability density (PDF), cumulative distribution (CDF), survivor, hazard, or cumulative hazard—can be expressed in terms of the other functions. *h*(*t*) and *S*(*t*)’s connection is through the basic identity:

$$S(t) = \exp\left(-\int_0^t h(u)du\right) \qquad 3$$

where *u* denotes all times at which we observe any failure event on the interval (0,*t*].[28] *S*(*t*) tells us the probability that a subject has *not* failed by time *t*. For our probability, we want to know whether the subject *has* failed by time *t*. This quantity is equal to the complement of *S*(*t*), $1 - S(t) \equiv F(t)$, *t*'s cumulative distribution function.

$$\Pr(y = 1) = F(t) = 1 - \exp\left(-\int_0^t h(u)du\right) \qquad 4$$

We now have our continuous-time Cox model expressed in terms of a subject's probability of failing by *t*.

B. Continuous Time to Discrete Time

We use "discrete time" to refer to situations in which a subject experiences failure sometime in a fixed interval, but we do not observe precisely when.[29] For instance, if we are recording BTSCS data on an annual basis, y_{ij} tells us our failure event happened sometime between the end of year $j - 1$ and the end of year *j*—($j - 1, j$]. Using a discrete-time model for BTSCS data tells us the probability of our event occurring sometime during year *j*. We compute *t*, the duration, by counting the number of years in which the subject has been at risk but not failed (or the number of years since the subject's last failure, if subjects can experience the failure event multiple times). Once at risk, the first year without a failure would be $t = 1$, the second year, $t = 2$, and so on.

[28] $\int_0^t h(u)du = H(t)$, the cumulative hazard function. The cumulative hazard's value in *t* will be the sum of all hazard values for every time point from 0 up through *t*.

[29] This particular situation is sometimes referred to "grouped duration data" or, sometimes, "interval-censored duration data" or "start-stop data" instead of discrete-time data. We refer to it as discrete time in this appendix for consistency with some authors' usage of the phrase, though truly discrete-time data and truly continuous data recorded at discrete intervals (= interval censored) are not synonyms, as we discuss in the main text.

TABLE 4. Data Structure for Two Subjects

CONTINUOUS TIME						DISCRETE/START-STOP TIME					
i	j_0	j	t_0	t	y	i	j_0	j	t_0	t	y
1	1989	1993	0	4	1	1	1989	1990	0	1	0
2	2004	2007	0	3	1	1	1990	1991	1	2	0
						1	1991	1992	2	3	0
						1	1992	1993	3	4	1
						2	2004	2005	0	1	0
						2	2005	2006	1	2	0
						2	2006	2007	2	3	1

Note: $y = 1$ if subject fails by t, 0 otherwise.

We acknowledge the shift to discrete time by modifying the interval over which we integrate the hazard. To explain by way of example: Table 4 shows data on two subjects, in a continuous-time format and a discrete-time format. Both formats use counting-process notation by adding a column for t_0 (Box-Steffensmeier and Jones 2004, 99–101). t_0 represents where the previous observation "left off" counting for t within a panel, and t continues to represent "up through the end of this time period," also within the panel. t's counting-process interpretation is consistent with our previous discussion of t in discrete time—it represents the probability of our event occurring sometime during year j (t), but after year j – 1 (t_0). For continuous time, the interval's starting point was implicitly $t_0 = 0$,[30] giving us our integration interval of $(t_0,t] = (0,t]$. For discrete-time data (e.g., annual), $t_0 = t - 1$. Our integration interval shifts for discrete-time data to reflect each observation's t_0. Equation 4, for discrete time, becomes:

$$\Pr(y_{ij} = 1) = 1 - \exp\left(-\int_{t-1}^{t} h(u)du\right) \tag{5}$$

If we insert equation 2 into equation 5 and simplify, we get:

$$\begin{gathered}\Pr(y_{ij} = 1) = 1 - \exp\left(-\int_{t-1}^{t} \left[h_0(u)\exp(\beta' X_{ij})\right]du\right) \\ \Pr(y_{ij} = 1) = 1 - \exp\left(-\exp(\beta' X_{ij})\int_{t-1}^{t} [h_0(u)]du\right)\end{gathered} \tag{6}$$

[30] $t_0 = 0$ implies no left truncation for either subject.

The final piece to recognize is that $\int_{t-1}^{t}[h_0(u)]du$ will return different values across *t*'s, but will be constant within a (*t* – 1, *t*] interval. Let $\alpha_t = \int_{t-1}^{t}[h_0(u)]du$, to reinforce that the integral's value is constant within each time interval. If we substitute α_t into equation 6 and then rearrange terms, we obtain:

$$\Pr(y_{ij} = 1) = 1 - \exp(-\exp(\beta' X_{ij})\,\alpha_t)$$
$$\Pr(y_{ij} = 1) = 1 - \exp\left(-\exp\left(\beta' X_{ij} + \ln(\alpha_t)\right)\right) \qquad 7$$

Moving α_t inside the parenthesis containing $\beta' X_{ij}$ makes clear that we are simply adding some constant—because the natural log of a constant is another constant—to the regression line (represented by $\beta' X_{ij}$). Adding a constant to the regression line changes the line's intercept by shifting it up or down. Since $\ln(\alpha_t)$ is indexed by *t*, the intercept shifts effectively give each *t* its own fixed effect. We can express fixed effects for *t* by including time dummies (denoted $\tau_t = \ln(\alpha_t)$ by BKT), giving us a temporal dependence correction. Equation 7 is identical to equation 1 if we added time dummies to the latter.

Appendix B:

Transition Probabilities: Background Information & Application

A. ORIGIN AND FORMULAE

Transition probabilities address the need for easier-to-understand interpretation techniques for hazard-based quantities. Additionally, our `mstatecox` package makes it easy to obtain confidence intervals around our transition probabilities in Stata, while the `mstate` package provides these abilities in R. Both have powerful implications for semi-parametric Cox models and their use.

The idea of obtaining *transition probabilities* from a duration model has its firmest roots in the multi-state duration model literature, but the quantity generalizes to any duration model. Usually, multi-state duration models are semi-parametric Cox models estimated with unique baseline hazards and unique covariate effects for every event within a process (see Metzger and Jones 2016 for an overview).[31] We focus on the semi-parametric variant, keeping with the paper's general focus. However, multi-state duration models can also be estimated parametrically or non-parametrically, meaning transition probabilities can also be calculated for any duration model—non-parametric, semi-parametric, or parametric.

Multi-state duration models begin by recognizing that we can view any process as being composed of a number of stages. Each stage is defined based on the event(s) subjects are at risk of experiencing—"risk sets," in duration parlance. Simple duration models, like basic L/P models, have two stages: (1) at risk of failing, which all subjects occupy to begin, and (2) failed, once subjects experience a failure event. A subject experiencing the event moves *from* one stage *to* another. We use the word "transitions" to denote this movement between stages, with its directed from-to pairings. The emphasis on stages opens up a new way of conceptualizing the output from duration models. Instead of thinking purely about *when* a subject experiences a transition, we can ask *which* stage it occupies at a given point in time.

[31] If needed, we can constrain different combinations of baseline hazards and/or the covariates' coefficients to be equal.

Defining transition probabilities begins with a familiar quantity: the hazard, often called a "transition intensity" in the multi-state literature and denoted $\alpha(t)$ instead of $h(t)$ because of the literature's roots in count models. We use a as a generic identifier for a subject's current stage ("from"), and b as the generic identifier for a subject's next potential stage ("to"). If we add transition-specific notation to the generic expression for a hazard, we obtain the transition-specific hazard—the instantaneous risk of transitioning from Stage a to Stage b at time t (Wreede, Fiocco, and Putter 2010, 262):

$$h_{a\to b}(t) \equiv \alpha_{a\to b}(t) = \lim_{\Delta t\to 0} \frac{\Pr(Z(t+\Delta t)=b|Z(t)=a)}{\Delta t} \tag{8}$$

where $Z(t)$ denotes the random process determining the stage's value in t (Stage a, Stage b, etc.).

Hazards and probabilities are similar, but not synonyms. The challenge becomes determining whether, and how, one of the quantities can be expressed in terms of the other, in such a way to yield transition probabilities. Gill and Johansen (1990, 1532–34) make the key connection in their work on product integrals. Drawing on work about counting processes, Gill and Johansen recognize the cumulative hazard function and the survivor function belong to a family of interval functions with well-known properties. They use these properties to map the cumulative hazard onto the survivor, and subsequently show this relation generalizes to more complex settings, where we have a process composed of many transitions, not just one (see also Aalen, Borgan, and Gjessing 2008, Appendix A; Andersen et al. 1993, 88–95).

As a consequence, we can aggregate every transition's cumulative hazard into an S x S matrix, $\mathbf{H}(t)$ (in the multi-state literature, $\mathbf{A}(t)$), where S is the number of stages within the process. After forming $\mathbf{H}(t)$ and applying Gill and Johansen's basic results, we obtain:

$$P(s,t) = \prod_{u\in(s,t]} (\mathbf{I} + \Delta\,\mathbf{H}(u)) \tag{9}$$

where u denotes all times at which we observe any transition within some time interval with start point s and end point t. The resultant quantities are now best described as transition probabilities from a Markovian process, contained in the $P(s,t)$ matrix. The matrix's quantities represent the probability of transitioning from each stage to every other stage within the time interval (s,t].[32] For example, element $P_{2,1}(s,t)$ would denote the probability of a subject transitioning from Stage 2 in time period s to Stage 1 by time period t. Metzger and Jones (2016, Appendix A) walk through a transition probability matrix calculation using a simple competing risks process.

As we mentioned above, equation 9 is contingent on the process having transition-specific hazards that are Markovian: a subject's next stage is conditional only on the subject's current stage. Practitioners can identify whether a hazard is Markovian via the transition-specific hazard expression. A Markovian transition-specific hazard's value does "not depend on any other aspect of the history, like states visited on the way, and the times of previous transitions (except to the extent that this information is reflected by the present state)" (Hougaard 2000, 143). If any covariates capture any of this information, the model is semi-Markovian in nature. If durations are recorded as gap time, the model is also semi-Markovian, since t is resetting *any* time a transition occurs, implying that t tells us something about the transition history. Whether a multi-state model is Markovian or semi-Markovian impacts whether or not we can calculate transition probabilities analytically.

B. Generating in R/Stata

Transition probabilities are a post-estimation quantity, like predicted probabilities from logit. We first estimate our Cox model, and then proceed to compute the transition probabilities. We elect to compute all our transition probabilities via simulation, for a few reasons. First, simulations can handle non-, semi-, and parametric duration models. Analytic expressions tend to be more rigid in their formulation, specific to a particular distribution or broader class of models. Second, our simulations do not make a Markov assumption about stages to estimate, whereas analytic expressions do. Third, our simulation setup is flexible and can accommodate a variety of process structures. This includes, but is not limited to, situations with recursive, repeated, competing, and/or sequential events. Analytic expressions

[32] The "within the time interval" in the interpretation stems from the order in which each term is multiplied together by the product integral.

have difficulty handling some of these alternatives, particularly recursive situations in which subjects can experience an event, and then once they experience a second event, become at risk of experiencing the first again. To obtain the transition probability analytically, researchers must be able to express every possible transition sequence a subject can take—a difficult proposition with recursiveness. Finally, a simulation approach makes generating confidence intervals simple, whereas analytic derivations are often quite complex. We simply look at the outputted transition probability from every simulation, and take the 2.5th and 97.5th percentile values for lower and upper CIs (for 95% CIs).

Wreede, Fiocco, and Putter's (2011) `mstate` package in `R` can calculate both analytic and simulated transition probabilities. The simulations are configured as a series of nested risk experiments. Our Stata package uses simulations exclusively, following the same setup as Wreede, Fiocco, and Putter (2011), which itself makes use of Dabrowska (1995).[33] Beyersmann, Allignol, and Schumacher (2011) also use the same setup to structure their book. We discuss our Stata package more extensively elsewhere (Metzger and Jones 2018), but the simulation broadly works as follows:

1. Decide on a number of subjects to move through the process, a starting time (*s*), a starting stage, and an end time (*t*) for all subjects.
2. For each subject:
    a. Set *s* to the current time and the starting stage as the current stage.
    b. Ensure the subject is at risk of a transition (with transition IDs notated *k*). If it is not (e.g., it is in an absorbing stage), move to the next subject.
    c. From the set of all observed failure times (*u*), randomly select a transition time larger than the current time, weighting each time's selection probability using the overall probability of any outward transition from the current stage at that time. Call this selected time t^*.

[33] Our package also has important additional functionality compared to its `R` counterpart: our package can handle covariates interacted with time, a common correction for proportional hazards violations, whereas `R-mstate` cannot (see Metzger and Jones 2018 for more details).

d. If $t^* > t$, subject stays in current stage until end of specified time interval t. Move to the next subject.

e. Otherwise, randomly select which transition the subject will experience, with each transition's selection probability being equal to $h_k(t^*)$.

f. Record t^* as the new current time and (2e)'s stage as the new current stage, for total time/clock time durations. Gap time/clock reset is the same, except that the new current time resets to 0.

g. Repeat (2b)–(2f) until the subject's $t^* > t$ or until the subject enters a stage with no outward transitions.

3. From the simulation results, count the number of subjects occupying each stage at each unit interval in $(s,t]$.
4. Loop over 1–3 for the desired number of simulation runs.

C. Dataset Structure

Structuring the data to account for a situation with two stages and two possible transitions requires (1) a stage identifier, recording whether a dyad is currently involved in a MID or not, as well as (2) a transition indicator, denoting when transitions between stages occur. Here, (2) amounts to a variable coded 1 for dyad-years in which MID initiation or termination occurs. We can then generate transition-specific covariates, allowing the covariates' effect to vary for MID onset and MID termination.[34] With this data structure, any standard statistical package can estimate this type of model, formally known as a multi-state model: estimate a Cox model, include the transition-specific covariates, and stratify the baseline hazards by stage.[35]

[34] For more on generating transition-specific covariates, as well as structuring data for more complex multi-state models, see Metzger and Jones (2016).

[35] Stratifying the baseline hazards by whether a dyad is currently experiencing a MID accounts for the different underlying rates of MID onset and termination. Because we only have two stages and two transitions, we can

D. Transition Probabilities For the MID Application

Interpreting the results of the multi-state model from the main text, which accounts for causal complexity and time-varying effects, is quite straightforward with the use of transition probabilities. Standard approaches to substantively interpreting Cox models focus on a covariate's effect on each transition in the process (e.g., its effect on event onset and its effect on event duration). This gives us insight into a covariate's effect on various events' occurrence, but we may also want to assess a covariate's effect overall, particularly in causally complex processes. To that end, we opt to interpret our Cox models using transition probabilities, which permit us to aggregate a covariate's effect across all transitions constituting the process. They are analogous, but not identical, to logit's predicted probabilities. A transition probability tells us the probability of a subject occupying a given stage by some point in time, given a set of starting conditions and a covariate profile of interest.

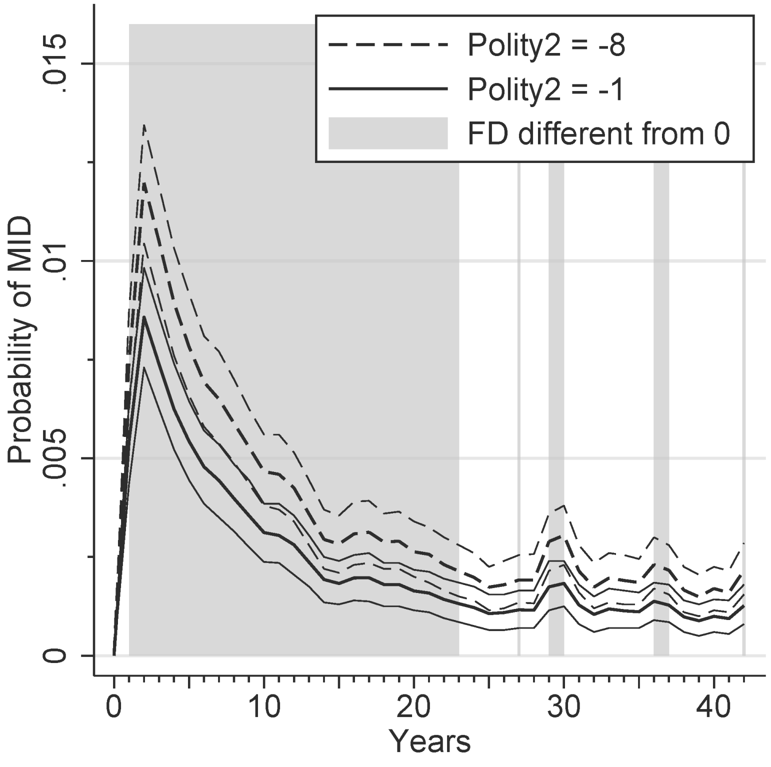
Using Table 3's model results, we simulate transition probabilities for two situations in which we vary the democracy levels in the dyad's least democratic state. In the first, we specify that the starting time is 0 and the starting stage is Peace, while setting the regime type measure to its 25th percentile. In the second, we keep the same starting stage and time, but set the regime type measure to its 75th percentile, holding all other covariates at their median values. Because Table 3's multi-state model captures both the determinants of MID onset (Peace → MID) and the determinants of MID duration (MID → Peace), our transition probabilities will assess regime type's overall effect on a dyad's militarized behavior.

Our resulting transition probabilities (Figure 4) depict the probability of a dyad being involved in a MID at different points across time, be it a new MID or an ongoing MID from previous years. The transition probabilities reflect *both* possibilities. Substantively, the transition probabilities continue to support much of the existing literature's findings: when the least democratic state in a dyad is a

stratify based on stage. If we had additional transitions among a larger group of stages, we would need to stratify on transition.

consolidated autocracy (Polity2 = -8), the probability of a dyad being involved in a MID is significantly higher than when the least democratic state in a dyad is closer to a democracy (Polity2 = -2). This effect exists consistently in the shorter- to medium-term only, though, from a year after the states' initial peace to 23 years out. The implication is more democratic dyads reduce the dyad's MID propensity soon after the dyad's peace initially begins, where the probability of being in a MID is high (relative to other *t* values). However, for longer-lasting spells of peace, democracy's pacific effect dissipates.

FIGURE 4. Transition Probabilities of MID Involvement



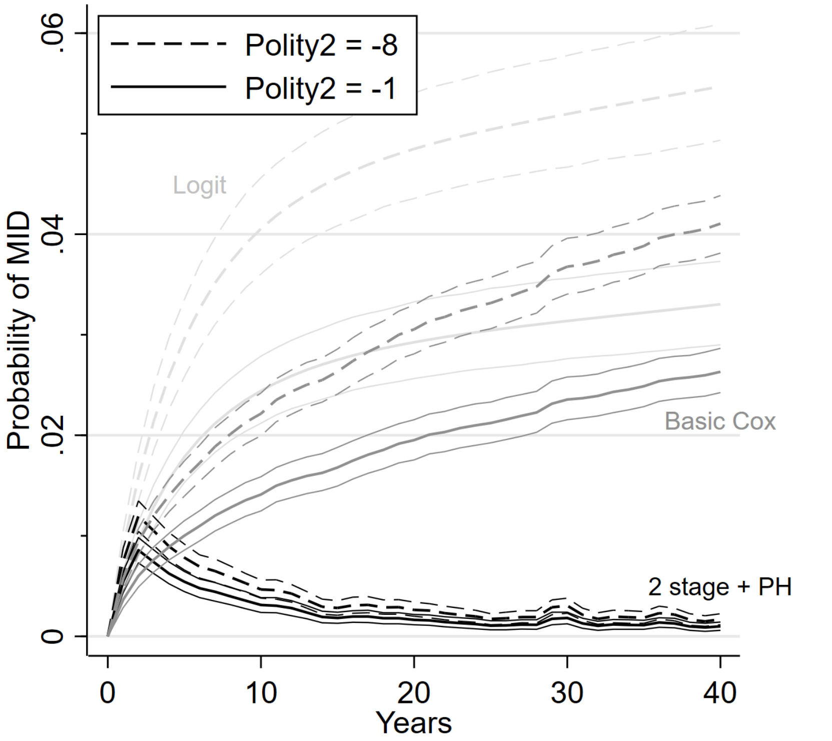
NOTE: thinner lines = 95% CIs
Level of democracy in dyad's least democratic state; 25th and 75th percentiles.

All other variables held at median.
1000 simulations, 20000 subjects per simulation

NOTE: Starting stage = Peace (no ongoing MID), starting time = 0.
FD = first difference.

To conclude, we compute the transition probabilities for the same scenario as above using Table 2's basic logit with time polynomials and its basic Cox model, and overlay them with Figure 4's transition probabilities (Figure 5). In name, all the transition probabilities represent the probability of two states being engaged in a MID at specific points in time.[36] Doing so makes our final point clear: the inferences we draw from a model that correctly adjusts for PH violations and causal complexity are vastly different in substantive meaning than those from an incorrect model with no PH correction or acknowledgement of causal complexity.

FIGURE 5. Comparing Transition Probs. of MID Involvement



NOTE: thinner lines = 95% CIs
Level of democracy in dyad's least democratic state; 25th and 75th percentiles.

All other variables held at median.
1000 simulations, 20000 subjects per simulation

NOTE: Transition probabilities generated with 1000 simulations, N = 20,000. Starting stage = Peace (no ongoing MID), starting time = 0.

[36] None of Table 2's models account for the return transition from MID to peace, meaning in practice, their transition probabilities represent the probability of experiencing a MID *onset* (vs. involvement) by specific points in time. This quantity is known as $F(t)$, the failure function, in the duration model literature. We do not contend this quantity is unhelpful, but we do note it derives from only one transition within the multi-transition process.

Appendix C:
Proportional Hazards Tests for Logit Models

A. Testing for PH

The procedure for testing for violations of the proportional hazards assumption in the context of BTSCS logit models is described by Carter and Signorino (2010).[37] Researchers must perform a series of likelihood-ratio tests. The restricted model in the tests is the base specification of the model, which assumes proportional hazards. The unrestricted specification(s) are similar to the base model, but include an interaction between a covariate and the three cubic polynomials. For each covariate in the model, it is necessary to estimate a unique model with time interactions. Once these models have been estimated, it is then possible to perform a series of likelihood-ratio tests of equivalence between the core restricted model and each of the unrestricted models. The result of these models indicates whether the interactions between each covariate and the cubic polynomials significantly improves model fit. If the test result is statistically significant, it indicates a likely violation of the PH assumption for that particular covariate. To correct for these violations, as with the Cox model, it is possible to include an interaction between the violating covariate and each of the terms in the cubic polynomial.

Table 5 presents the results of tests for violations of the PH assumption for the logit model in Table 2 of the main text. Each row contains the result of a likelihood-ratio test of the core restricted model, and an unrestricted specification of the model including an interaction between the covariate and each of the three time variables. Table 5 indicates that each of the covariates in the model likely violates the PH assumption, as the test statistic for each likelihood-ratio test is significant at the $p < 0.000$ level. To correct for these violations, it is necessary to include interactions between each of the covariates in the model, and each of the time terms. This result is the same as that attained from tests of Schoenfeld residuals following estimation with a Cox model.

[37] As we mentioned in the main text, logit technically assumes proportional odds, not proportional hazards.

TABLE 5. Proportional Hazards Tests for Logit Specification

	χ^2	*p*-value
Allies	26.06	0.000
Capability Ratio (ln)	45.57	0.000
Economic Interdependence (Low)	47.75	0.000
Contiguity	93.90	0.000
Distance (ln)	19.17	0.000
Major Power Dyad	18.76	0.000
Democracy (Low)	46.97	0.000
Joint IGOs	89.38	0.000

Test statistics result from a likelihood-ratio test of equivalence between the base restricted model, and an unrestricted variant with time interactions for each covariate.

B. LR Test's Performance

We performed a small set of simulations to assess the statistical power of the LR test, for PH detection purposes. In the simulation, we include a single covariate, $x \sim N(0,1)$. We vary four characteristics:

- *N*: {100, 250, 500}
- Baseline hazard: $\{t, \ln(t), |0.4t^2 - 0.16t - 1.3|\}$ ⇒ also used as function for PH violations
- x's time-varying effect: {-0.2, -0.05, 0, 0.05, 0.2} ⇒ coefficient for x*(time funct) interaction
- x's main effect: {-0.5, -0.2, -0.05, 0, 0.05, 0.2, 0.5} ⇒ coefficient for x constituent term

Forming the combination of all these characteristics produces 315 scenarios. We perform 1000 simulation draws for each scenario. We use the same procedure we describe in Appendix G to generate the data: we use `survsim`, and then coerce the resultant durations into start-stop format. For logit, we first run a model with time polynomials, and then run a second model in which the time polynomials are interacted with *x*. We subsequently use the LR test, described in the previous section, to assess whether evidence of a PH violation exists. For comparison, we also run a Cox model and run Therneau and Grambsch's standard PH test for *t*, ln(*t*), rank(*t*), and the Kaplan-Meier. These time transformations are standard for Therneau and Grambsch's test, and come included as standard options in R's and Stata's canned PH test routine.

Our key quantities of interest are the LR test's *p*-value, and the Cox PH test's *p*-values (of which there will be four, one for each time transform). When *x*'s time-varying effect is non-zero, a PH violation exists. The tests should reject the null hypothesis of no violation. The conventional threshold for power calculations is 80% (Aberson 2019, 8), meaning the test's *p*-values should be less than 0.05 in at least 800 of the 1000 of the simulation draws. These scenarios give us a sense of how well the tests detect violations, when they exist. When *x*'s time-varying effect is 0, no PH violation exists. Accordingly, the tests should fail to reject the null. In practical terms, for the no effect scenarios, the various tests' *p*-values should be less than 0.05 in 5% of the simulation draws, speaking to the tests' statistical size. We include the full logs as part of our replication package.

Our simulation results show that both PH tests are poorly powered. In scenarios where a PH violation exists, both tests often fail to detect it more than 80% of the time (Type II error). Others have also pointed out the Cox PH tests' poor power (Austin 2018), but the logit's poor power in PH contexts is less known. Conversely, Cox's test does better for the null scenarios (Type I error)—where *x*'s time-varying effect is zero—either performing on par or better than logit's LR test.

Appendix D:
Hall and Ura

A. Main Analysis

To show how transition probabilities can help improve our interpretation of Cox model results, we re-examine Hall and Ura's (2015) study of judicial invalidation of major federal laws, purposely changing nothing about their models except the estimator, to illustrate the similarities between L/P and the Cox and the general virtues of transition probabilities. We assume the data exhibit no causal complexity in this portion of our analysis, the same as Hall and Ura.[38]

Hall and Ura argue the judiciary is more likely to invalidate significant legislation if the current legislature demonstrates little support for the legislation. To test this claim, they construct a dataset of US statutes from 1949–2008. The unit of analysis is the statute-year, and the dependent variable is a dichotomous indicator of whether the US Supreme Court invalidates (either fully or partially) the statute in a particular year. Hall and Ura employ three different independent variables, each of which attempts to measure the current pivotal legislator's degree of support for the statute. Each variable captures the likelihood that the pivotal legislative actor would vote in favor of the law, but varies who is defined as being pivotal. The Floor Median Model focuses on each chamber's median member as well as the president, the Senate Filibuster Model includes filibuster actors, and the Party Gatekeeping Model includes the majority party's median member in both the House and the Senate (2015, 823).

To model the relationship between legislative support for a statute and the likelihood of judicial invalidation, Hall and Ura estimate a series of logit models. As part of their model specification, they include a counter of the number of years since a statute was passed *or* since the statute was last partially invalidated (i.e., a gap-time duration),[39] along with cubic polynomials of the counter, as suggested by Carter and Signorino (2010). In explaining their modeling strategy, Hall and Ura state—*correctly,* we

[38] In the next section, we use Cox models to better model the causal complexity in these data. Doing so leads to different substantive conclusions.

[39] We return to the issue of multiple invalidations in the next section, where we show how failing to acknowledge the possibility of repeated invalidations leads to different substantive conclusions.

would argue—that their "approach is functionally equivalent to a traditional duration analysis *and offers clearer interpretation*" (2015, 824, emphasis added).

1. Model Estimates

Table 6 replicates all three of Hall and Ura's logit models, and re-estimates each as a Cox model.[40] As Table 6's coefficient estimates indicate, both the logit and Cox estimation strategies yield similar inferences. Regardless of how legislative support is operationalized, as support for a statute increases, the Supreme Court becomes less likely to invalidate (partially or fully) that statute. However, focusing solely on coefficient signs limits researchers' ability to substantively interpret their results. In the main text of their manuscript, Hall and Ura use predicted probabilities and marginal effects to substantively interpret their logit models' estimates.

2. Interpretation via Transition Probabilities

Estimating transition probabilities from duration models provides more intuitive substantive interpretations than most current duration post-estimation strategies, and makes estimating confidence intervals straightforward. To highlight these features, we estimate transition probabilities using Table 6's Floor Median model. Specifically, we simulate transition probabilities for two situations. In the first, we specify that the starting time is 0 and the starting stage is stage 1, meaning the law has just passed. In the second, we keep the same starting stage, but specify the starting time as 10, meaning the law was passed 10 years ago. The debate about invalidating the Affordable Care Act (enacted in 2010) underscores why a scenario with a non-zero starting time might be useful. We estimate these transition

[40] There are several minor differences in logit coefficient estimates from Hall and Ura's study, due to a difference in the construction of the gap-time counter. Following a judicial invalidation, Hall and Ura reset time to 1 and begin counting anew, rather than 0. Employing this method, we can perfectly replicate the reported results. However, we adopt the more standard approach of resetting time to 0 for our analysis.

probabilities for two scenarios: one with high legislative support for the law and one with low legislative support.

TABLE 6. Judicial Review – Comparing Logit and Cox Estimates

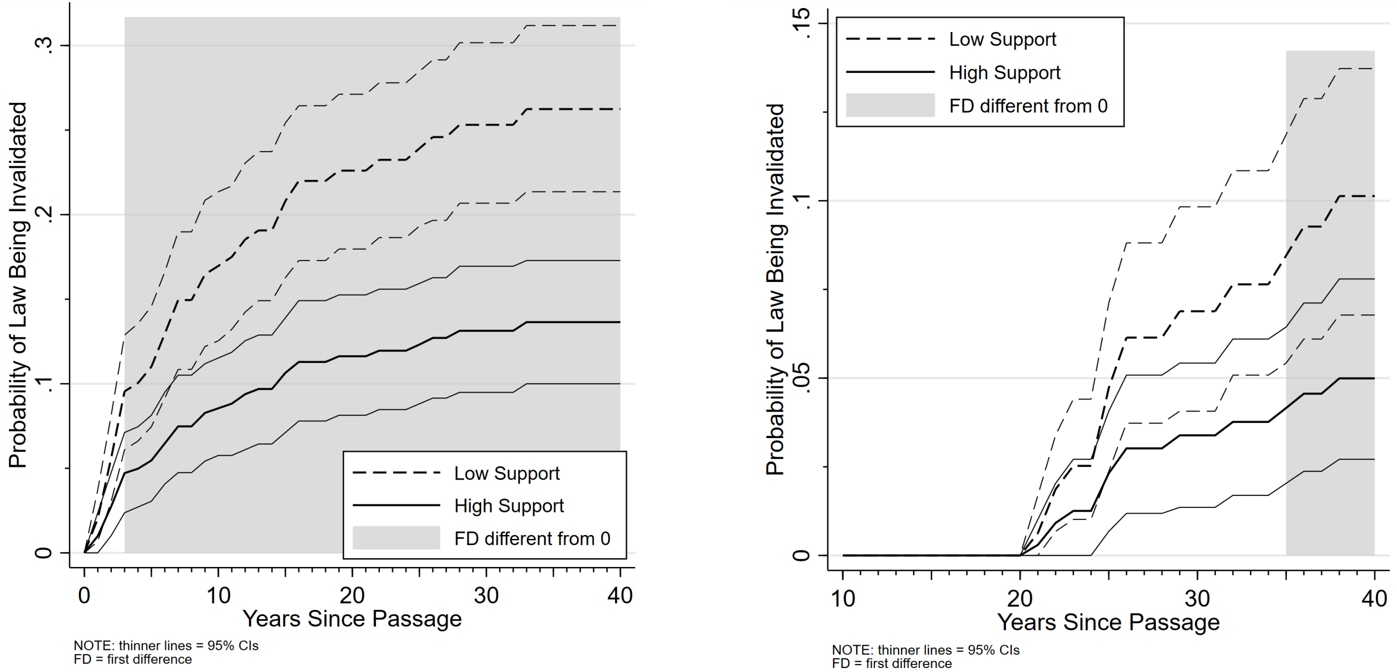
	Floor Median		Senate Filibuster		Party Gatekeeping	
	Logit	Cox	Logit	Cox	Logit	Cox
Majority Support	-1.45** (0.52)	-1.44** (0.52)	-1.12* (0.50)	-1.11* (0.49)	-1.29** (0.50)	-1.28** (0.49)
Constant	-2.51** (0.58)	--	-2.88** (0.53)	--	-2.79** (0.51)	--
Log-Likelihood	-273.6	-264.4	-274.8	-265.5	-273.8	-264.6

† = $p \leq 0.10$, * = $p \leq 0.05$, ** = $p \leq 0.01$, two-tailed tests. Cox coefficients reported in terms of hazards. Robust standard errors reported for both models. Partial log-likelihood estimates reported for Cox models. Logit models also include a counter for years since a statute passed *or* was last invalidated, along with quadratic and cubic terms.

Figure 6 plots the simulated transition probabilities. There are two possible stages a law can occupy in a given year: not yet invalidated and already invalidated. Figure 6's values reflect the probability that a law *has been* declared invalid by that particular year. For instance, the transition probability at $t = 5$ reflects the probability of the law being invalidated in year 5, but also the probability that the law was invalidated in $t = 4$, $t = 3$… $t = 1$. Thus, the probabilities reflect the overall likelihood that a law has been declared invalid by each year, regardless of *when* it was declared invalid, which is different from the probability of invalidation happening in any given year. Notably, invalidation is not absorbing.[41] Once a law has been invalidated in a given year, it remains at risk of being invalidated in the future, since statutes may be invalidated multiple times, potentially as different components of a law are subject to scrutiny.

[41] This is a property of the substantive example, not transition probabilities themselves, as transition probabilities can readily accommodate absorbing states.

FIGURE 6. Transition Probabilities of Judicial Review



(a) Starting time = 0

(b) Starting time = 10

NOTE: Transition probabilities generated with 1000 simulations, $N = 295$. Starting stage = law valid (both panels).

Figure 6 depicts several interesting results that demonstrate transition probabilities' utility. First, focusing on Figure 6a, there is a relatively high chance that significant legislation will be invalidated. Regardless of the level of legislative support, there is roughly a 5–15% chance that a major piece of legislation will be ruled invalid within 5 years of its passage. Starting with the third year after passage, laws with low legislative support have a significantly higher probability of being invalidated than laws with high legislative support, indicated by the first difference being statistically distinguishable from zero (shaded regions). Focusing only on the probability of invalidation happening in any given year—which may be relatively low—risks obscuring these larger points about invalidation across time. Notice also how the figure's confidence intervals make it easy to evaluate whether the two scenarios' transition probabilities are statistically different from one another.

Second, and related to the prior point, transition probabilities are readily interpretable and in an intuitive scale. This facilitates comparisons between different values (e.g., are the chances of invalidation higher when legislative support goes down?), but also absolute inferences. For instance, there is a 25.3% chance that major legislation will be invalidated within 30 years since the law was passed when there is low legislative support (Figure 6a). This value clearly indicates preserving major legislation is far from guaranteed, especially when such legislation lacks popular support.

Third, transition probabilities give us more flexibility to specify additional situations of interest. Figure 6b shows that, if a law has been on the books for 10 years, it will have a 4.7% chance of being invalidated in the presence of low legislative support at the 25-year mark, whereas it will have a 2.3% chance of being invalidated in the presence of high legislative support at the 25-year mark. However, these differences only become statistically distinguishable from one another 35 years after the law passes, suggesting that *if* laws survive to the 10-year mark, there is a small buffer zone from 10 to 35 years in which the degree of legislative support has no significant effect on the law being invalidated. This sort of inference, where our starting time for the simulations is not 0, are not possible with our current duration model interpretation strategies.

Finally, this application showcases L/P's limited ability to model causal complexity. Hall and Ura's statutes can be fully *or* partially invalidated. Because of partial invalidation, we can observe multiple invalidations for a single statute. In Hall and Ura's dataset, 37 laws are invalidated once, 9 laws are invalidated twice, 4 laws are invalidated three times, and 1 law is invalidated four times. However, L/P models have trouble easily accounting for this type of recurrent event. In the analysis above, the time counter resets to 0 each time a law is partially invalidated, a strategy referred to as either gap time or clock-reset time in the duration literature. The rationale is time since the *previous* event may affect the probability of *subsequent* events. However, a gap-time coding does not acknowledge an additional implication of event recurrence: event dependence. Statutes invalidated once could be somewhat more (or less) likely to be invalidated a second time, twice-invalidated statutes may be even more likely to experience a third invalidation, and so on.[42] Instead, basic L/P models view all invalidations as identical by pooling across each of these situations, unless practitioners take specific steps to the contrary. By

[42] Among duration models, repeated events models specialize in recognizing event dependence of this nature. See Box-Steffensmeier and Zorn (2002).

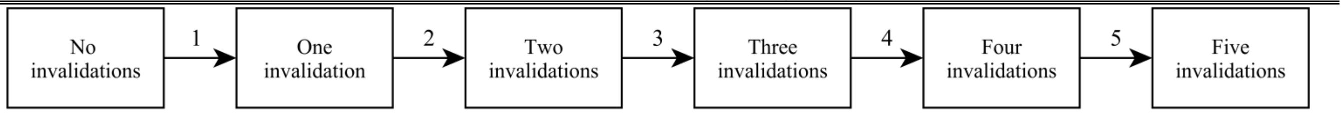
contrast, Cox models have an elegant and straightforward solution to the potential event dependence implied by recurrent events—stratification, which is the matter we turn to next.

B. Causal Complexity: Repeated Events

As we mentioned above, statutes can be invalidated multiple times in Hall and Ura's dataset. As we also mentioned above, our analysis treats each successive invalidation as identical, the same as any basic L/P analysis does. Specifically, we assume the underlying rate at which each invalidation occurs is the same (equivalent baseline hazards), and we assume legislative support's effect for every invalidation is the same (equivalent covariate effects). If these assumptions are incorrect, our model estimates will be incorrect, yielding erroneous inferences (Metzger and Jones 2016).

Whether or not these assumptions are valid is testable using Cox models, as Metzger and Jones discuss (2016, 469–72). We apply their procedure to assess these two assumptions. We find one of the assumptions is partially invalid, while the other is entirely valid. Importantly, when we correct for the partial assumption violation, the substantive conclusions we draw from our transition probability graphs are different from the previous section's graphs. Legislative support has a significant effect on the probability of a statute being invalidated starting in year 6 (vs. the last section's year 3).

FIGURE 7. Hall and Ura: Full Stage Diagram



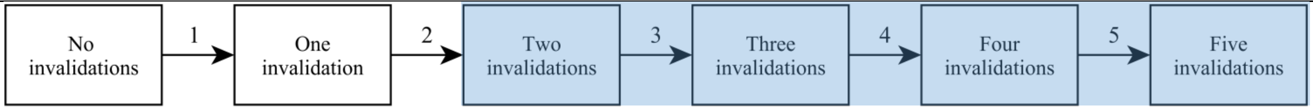
	No invalidations	One invalidation	Two invalidations	Three invalidations	Four invalidations	Five invalidations
No. of statutes	260	34	9	3	1	0
No. of events	35	9	4	1	0	-

NOTE: "event" = "exiting transitions from the stage." Number of exiting transitions ≠ statutes in next stage because two statutes are partially invalidated in 2008, the last year of Hall and Ura's sample.

We begin by estimating a model in which each invalidation (represented by an arrow in Figure 7) has a separate baseline hazard rate and an invalidation-specific effect for legislative support. However, this model's results are nonsensical, unsurprisingly (e.g., some of the coefficients do not estimate). There are very few transitions between the last four stages, and Cox models perform poorly in the presence of exceptionally few transition events, similar to L/P models (see Metzger and Jones 2016, Online Appendix

L). There is not enough information being fed into the Cox model to obtain all the parameter estimates we have specified.

FIGURE 8. Hall and Ura: Constrained Stage Diagram ($h_0 + \beta s$)



No. of statutes	260	34	9	3	1	0
No. of events	35	9	4	1	0	-

NOTE: Baseline hazard and legislative support's effect constrained to be equal for shaded transition arrows. "event" = "exiting transitions from the stage." Number of exiting transitions ≠ statutes in next stage because two statutes are partially invalidated in 2008, the last year of Hall and Ura's sample.

TABLE 7. Judicial Review – Cox Estimates, Constrained Stages

	Floor Median				
	Transition ID #1	Transition ID #2	Transition ID #3	Transition ID #4	Transition ID #5
Majority Support	-1.25*	0.27	-4.08*	-4.08*	-4.08*
	(0.56)	(1.70)	(1.62)	(1.62)	(1.62)
Partial LLH	-222.3				

† = $p \leq 0.10$, * = $p \leq 0.05$, ** = $p \leq 0.01$, two-tailed tests. Robust standard errors reported. LLH = log-likelihood. Shaded cells = parameters constrained to be equal. Shaded model headers = transitions' baseline hazards constrained to be equal.

We constrain both the baseline hazards and covariate effects for transitions 3–5 to be equal, reducing the number of parameters the Cox model needs to estimate. We then estimate the model again, and this time, we do get estimates for all our parameters successfully (Table 7). We subsequently use this model to test the two assumptions we mentioned previously. With the constraints we have imposed on the model, the assumptions now specifically amount to the following:

1. Equivalent baseline hazards: Does transition 1's baseline hazard equal transition 2's baseline hazard, and does transition 2's baseline hazard equal transitions 3–5's baseline hazard? Substantively, loosely put, when the statute has a 0% chance of being supported by the legislature, is Pr(first invalidation) equal to Pr(second invalidation) equal to Pr(three or more invalidations)? $\left(h_{\text{tr1}_0}(t) = h_{\text{tr2}_0}(t) = h_{\text{trs3}-5_0}(t)\right)$

2. Equivalent covariate effects: Does the legislative support coefficient for transition 1 equal its coefficient for transition 2, and does the legislative support coefficient for transition 2 equal its coefficient for transitions 3–5? Substantively, and again loosely put, is legislative support's effect on Pr(first invalidation) the same as its effect on Pr(second invalidation) the same as its effect on Pr(three or more invalidations)? $\left(\beta_{\mathrm{tr1(MAJ\,SUP)}} = \beta_{\mathrm{tr2(MAJ\,SUP)}} = \beta_{\mathrm{trs3-5(MAJ\,SUP)}}\right)$

We perform the tests described by Metzger and Jones (2016, 469–72) using Table 7's model. We find one of the assumptions is partially invalid, while the other is entirely invalid:[43]

1. Baseline hazards: Partially invalid assumption. The baseline hazards of the first invalidation and second invalidation are not statistically different from one another (valid), and can be collapsed into one. However, the baseline hazards of one or two invalidations compared to three or more invalidations are statistically different from one another (invalid), and should therefore be estimated separately. As a consequence, we can estimate only two baseline hazards: one for the first and second invalidation {transitions 1–2} and one for three or more invalidations {transitions 3–5}.
2. Covariate effects: Invalid. Our Wald tests indicate Table 7's $\beta_{\mathrm{tr1(MAJ\,SUP)}}$ is statistically distinguishable from $\beta_{\mathrm{tr2(MAJ\,SUP)}}$, which in turn is statistically distinguishable from $\beta_{\mathrm{trs3-5(MAJ\,SUP)}}$. The number of prior invalidations conditions legislative support's effect on additional invalidations.[44] We should estimate three different effects for legislative support.

Table 8 contains the final Cox model implied by our tests. It contains three estimated covariate effect and two unique baseline hazards. The difference between Table 8's model and the previous section's model (Table 6) is twofold: (1) permitting the first two and last three transitions to have different underlying rates of occurrence (baseline hazards) and (2) permitting legislative support's effect to differ, depending on how many previous invalidations a statue has experienced. The substantive

[43] The specific tests and their results are included as a do-file in our replication package.

[44] The Wald tests do suggest $\beta_{\mathrm{tr1(MAJ\,SUP)}}$ and $\beta_{\mathrm{trs3-5(MAJ\,SUP)}}$ are indistinguishable from one another. However, from a substantive perspective, it is unclear why legislative support's effect on the first invalidation should be equal to its effect on three or more invalidations. As a result, we leave these two effects separate.

conclusion we would draw from Table 8's model estimates alone is the same as from Table 6: statutes with higher levels of legislative support are less likely to be invalidated.

TABLE 8. Judicial Review – Cox Estimates, Final Model

	Floor Median				
	Transition ID #1	Transition ID #2	Transition ID #3	Transition ID #4	Transition ID #5
Majority Support	-1.29*	-0.01	-4.08*	-4.08*	-4.08*
	(0.56)	(0.72)	(1.62)	(1.62)	(1.62)
Partial LLH	-243.1				

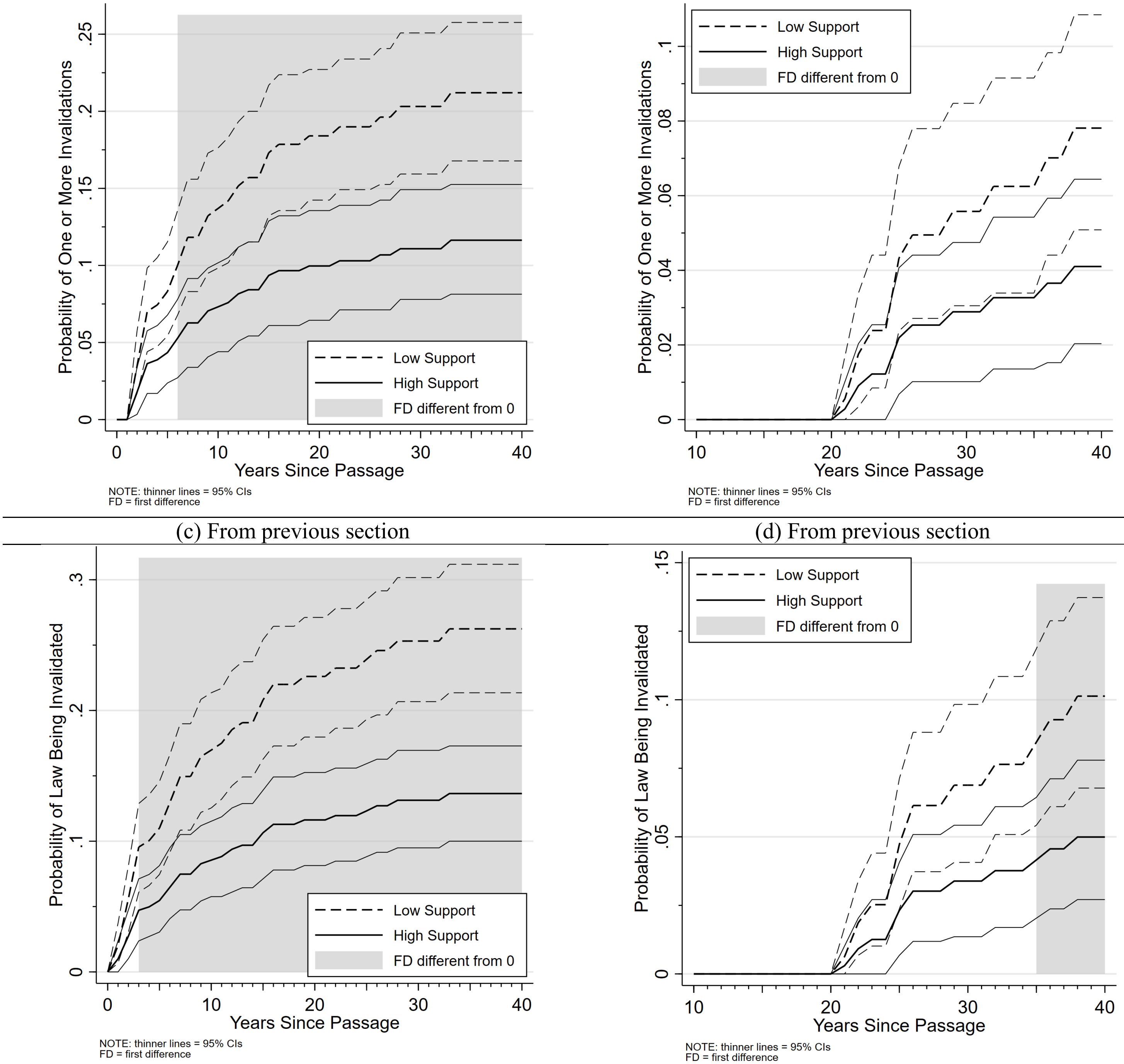
† = $p \leq 0.10$, * = $p \leq 0.05$, ** = $p \leq 0.01$, two-tailed tests. Coefficients reported in terms of hazards. Robust standard errors reported. LLH = log-likelihood. Shaded cells = parameters constrained to be equal. Shaded model headers = transitions' baseline hazards constrained to be equal.

We generate transition probabilities from Table 8's model next, to substantively interpret the model results. We create the graphs in such a way that they are directly comparable to the previous section's analysis. These transition probabilities (Figure 9's top row) paint a very different picture than the previous section's analysis (reproduced as Figure 9's bottom row). The previous section's Panel A (Figure 9c) showed the probability of invalidation for high-support statutes was statistically different from those with low legislative support after the 3-year mark, given the law had just been passed. However, Figure 9a shows this is no longer the case. Once we properly acknowledge the underlying rates at which invalidation events occur, the degree of legislative support no longer affects whether and when a statute is invalidated until the 6-year mark—the equivalent of 3 new elected Congresses vs. the 2 elected Congresses implied by the previous section's analysis. The conclusions we draw from Figure 9b are also different from the previous section: 10 years after the statute's passage, legislative support has no effect on the probability of invalidation at all.

This reanalysis demonstrates how we may otherwise overlook inferences about a process by treating all the process' events as identical, plus the importance of generating predicted quantities. Table 8's results suggested our substantive conclusions did not change much if we acknowledged the recurrent nature of statute invalidation. We only detected legislative support's different effect—and the different conclusions it implies—once we generated transition probabilities using Table 8's results. The model results alone were insufficient.

FIGURE 9. Final Model, Transition Probabilities of Judicial Review

STARTING TIME = 0 | STARTING TIME = 10


(a) Table 8's model

(b) Table 8's model

(c) From previous section

(d) From previous section

NOTE: thinner lines = 95% CIs
FD = first difference

NOTE: thinner lines = 95% CIs
FD = first difference

NOTE: thinner lines = 95% CIs
FD = first difference

NOTE: thinner lines = 95% CIs
FD = first difference

NOTE: Transition probabilities generated with 1000 simulations, $N = 295$. Starting stage = No invalidations (all panels)

Appendix E:
Cox Model Disadvantages

Any econometric model has advantages and disadvantages. The Cox model is no exception. Our discussion in the main text focused on the model's advantages, particularly in a BTSCS context. The Cox's lack of an easy-to-interpret post-estimation quantity often comes up as its biggest disadvantage, usually implicitly. Transition probabilities address this issue and render it moot. Beyond this, the Cox model has at least three principal "disadvantages":[45]

1. Prediction/Forecasting: Cox models' semi-parametric estimation trades off information about the hazard's functional form for the ability to make inferences about covariate effects. The hazard's functional form is useful for generating predictions, because it specifies how the hazard behaves for any time point where we do not observe a failure. Since the Cox makes no functional form assumptions, we do not know this information. Unsurprisingly, then, the Cox model does not do well at prediction/forecasting out of sample (Box-Steffensmeier and Jones 2004, 86). Relatedly, Cox models run a higher risk than normal of overfitting the data, which is true of any non- or semi-parametric estimation method (Box-Steffensmeier and Jones 2004, 89).
2. Computational Speed: The canonical Cox model is estimated using partial likelihood. Partial likelihoods are sequentially calculated: informally, when the first subject fails at time t_1, we use information about everyone else that *could* have also failed at t_1; when the second subject fails at t_2, we use information everyone else that *could* have also failed at t_2 (which will be everyone except subject 1, since s/he already failed); and so on for every subject with an observed failure. For every additional subject that fails as time passes, the number of subjects still able to fail ("at risk," in duration parlance) gets increasingly smaller. However, precisely how many subjects are at risk at a later time point is not known until the calculations for earlier failure times are complete. The sequentiality means parallel processing can be difficult to bring to bear when

[45] Box-Steffensmeier and Jones (2004, chap. 6) provide a longer discussion about when to use parametric vs. semi-parametric duration models. Their broad comments about when parametric duration models are preferable to Cox models apply to L/P vs. Cox as well, since both parametric duration models and L/P parametrize the baseline hazard.

estimating Cox models.[46] As a result, Cox models may take slightly longer to estimate than standard L/Ps, particularly for large datasets.[47] In practice, we have not noticed a prohibitively large difference in Stata, aside from fn. 47's situations.

3. Scope Conditions: The Cox transition probability estimates behave well in the situations political scientists are most likely to encounter (i.e., coerced start-stop data, a.k.a. interval-censored data). However, there are some situations where the Cox transition probabilities' behavior is less reliable. Others have established that the Cox does poorly at recovering coefficient estimates when many subjects experience the event at the same recorded time ("ties") (Hertz-Picciotto and Rockhill 1997). We discuss these rough scope conditions in Appendix I. In situations where the scope conditions may not be met, practitioners should weigh the merits of the Cox vs. logit much more deliberately and cautiously.

[46] E.g., for Stata: http://blog.stata.com/2010/11/11/statamp-having-fun-with-millions/.

[47] For example, the main text's MID example has 465,997 observations. In Stata 14.2 MP, using two cores, the logit with cubic polynomials takes ~5 seconds to estimate, while Table 3's multi-state Cox model with manual PH corrections took the longest at ~14.6 seconds. The respective models take longer to estimate if you use factor notation (e.g., `c._t##c._t##c._t` for the logit's cubic polynomials) or `stcox`'s `tvc()/texp()` options to avoid manually generating time interactions. For the `stcox` case, the estimation time is *significantly* larger with `tvc()/texp()` (~740.7 seconds vs. the ~14.6 we mentioned earlier). When possible, we recommend `stsplit`ting the data (if needed) and manually generating time interactions for any PH violations (related to this point, see Jin and Boehmke (2017)).

Appendix F:
Cox Model 101

A. Exposition

To see how Cox models can be useful in a BTSCS setting, understanding the model's basics is helpful.[48] A Cox model is a type of duration model, of which there are several different types.[49] As a general class of models, duration models are concerned with lengths of time as a dependent variable. Duration models are useful when investigating questions about how long before a subject experiences some event of interest. We count the number of periods the subject "survives" before experiencing our event; the resultant quantity is t, our duration of interest and duration models' dependent variable. Duration models' focus on time is what corrects for any temporal dependence in the data.

Duration models can be principally formulated (and their subsequent log-likelihoods expressed) in one of two ways, though others do exist. The first is in terms of x's effect on the duration (known as an "accelerated failure time" [AFT] metric), and the second is in terms of x's effect on the hazard of the event's occurrence (known as a "proportional hazard" [PH] metric). We mention the different metrics only to stress they are two different ways *to speak about the same thing*: our event of interest, and what factors increase or decrease the likelihood of its occurrence.

Cox models can only be expressed in a PH metric, which will have downstream implications for substantive interpretation. The generic form for any (PH-based) hazard is:

$$h(t \mid X) = h_0(t)\exp({\beta_{PH}}'X) \tag{10}$$

$h(t)$ represents the hazard function, also known as the hazard rate—the instantaneous rate at which the event of interest occurs at t.[50] Equation (10)'s right-hand side makes clear why this metric is named "proportional hazard." $h_0(t)$ is the baseline hazard, akin to OLS' intercept term. It represents the event's hazard of occurring when there are either no covariates in the model or when all of the covariates are set

[48] We encourage practitioners interested in using the Cox to read up further. For good introductions, see Singer and Willett (2003, chaps. 14–15) and Box-Steffensmeier and Jones (2004, chap. 4). For a more advanced discussion, see Therneau and Grambsch (2000, chap. 3).

[49] For a refresher, see Jones and Metzger's (2019) supplemental appendix A for a brief overview, and Box-Steffensmeier and Jones (2004) for a more thorough treatment.

[50] In continuous-time duration models. For discrete-time duration models, $h(t)$ represents a conditional probability, not a rate.

to zero. The covariates, appearing in the second term, shrink or enlarge $h_0(t)$'s value, depending on whether $\exp(\beta_{PH}{}'X)$ is greater than 1 (enlarging $h_0(t)$) or less than 1 (shrinking). Taken together: $h(t \mid X)$ is expressed as a scaled-up or scaled-down value of $h_0(t)$; $h(t \mid X)$ is expressed as a *proportion* of $h_0(t)$.

Cox models are popular because of the way in which they treat $h_0(t)$. Normally, to estimate a duration model with covariates, we must make some assumption about $h_0(t)$'s functional form—does the event's occurrence become more likely as time passes? Less likely? Does it become more likely and then less? These are parametric duration models—we impose a specific functional form for $h_0(t)$, allowing us to obtain our β estimates using maximum likelihood. By contrast, the Cox model is *semi*-parametric: it makes no $h_0(t)$ functional form assumption. The canonical Cox model uses partial likelihood to obtain its β estimates, and partial likelihood does not require us to make any assumptions about $h_0(t)$'s functional form. Importantly, the Cox's partial-likelihood estimates of β will still be unbiased. It is hard to overstate the significance of this insight, and the Cox model more broadly, to the duration modeling community. To give a sense of the model's import: Sir David Cox's (1972) paper introducing the model has over 51000 citations on Google Scholar as of January 2020.

B. Common Quantities

The four most common Cox model post-estimation quantities are:[51]

- Hazard ratio: how many *times* larger (or smaller) the hazard's value becomes for a one-unit increase in x, holding all else constant (Box-Steffensmeier and Jones 2004, 50; Mills 2011, 94). Hazard ratios are the only quantity with reported p-values by default for both R's `survival` package and Stata's built-in duration commands.
- Percent change in hazard rate: by what *percentage* the hazard's value changes for an increase (usually by one unit) in x's value, holding all else constant (Box-Steffensmeier and Jones 2004, 60; Mills 2011, 95). R's `survival` package has no function to calculate percent

[51] Jones and Metzger (2019) provide a more thorough exposition of various duration model interpretation techniques, complete with examples and discussions of the various techniques' merits. They also discuss different existing user-written packages to compute duration post-estimation quantities in both R and Stata.

change and its corresponding CIs. Instead, R users must calculate percent change using the `simPH` package's `coxsimLinear` to obtain CIs easily. In Stata, users must manually calculate percent change using `nlcom`, which will also return a *p*-value/CIs.

- Cumulative hazard (function): how many failures we would expect to see in the time interval 0 to t, given a particular set of covariate values ("covariate profile")[52] (Cleves et al. 2010, 13–15; Singer and Willett 2003, 488–91); adds together $h(t)$'s value for every t in a given range: cumulative hazard $= H(t) = \int_0^t h(u)du$. R's `survival` package can estimate both $H(t)$ and confidence intervals. Stata's built-in duration commands can generate $H(t)$ from a Cox model, but cannot generate CIs. The user-written `survci` command can do both.
- Finally, the survivor function is the sole non-hazard-based quantity commonly used to interpret Cox model results. The survivor's value tells us how many subjects' failure times (denoted T in the formula) are above a specific time (usually denoted t); defined formally as survivor function = $S(t) = \Pr(T \geq t)$.[52] Similar to $H(t)$, R's `survival` package can estimate both $S(t)$ and confidence intervals. Also similar to $H(t)$, Stata's built-in duration commands can estimate a Cox model's $S(t)$, but cannot generate CIs; the user-written `survci` command can do both.

C. Regarding Transition Probabilities

Standard approaches to substantive interpretation of Cox models focus on modeling the hazard of some event occurring (the hazard rate), or alternatively, the probability of that event *not* occurring (the survival probability) (Jones and Metzger 2019). While useful for gaining insights into the risk of a single event occurring, in many instances, such as MIDs, we may be equally, if not more, interested in the

[52] Unlike the first two quantities, researchers would need to compute the cumulative hazard ($H(t)$) for multiple covariate profiles, in order to effectively demonstrate how much $H(t)$ changes by for some change in x. Researchers would also need confidence intervals around the different covariate profiles' $H(t)$s, to see whether the $H(t)$ values are statistically different from one another, indicated by no CI overlap (though overlapping CIs does not necessarily mean statistical insignificance; see Austin and Hux 2002). The same is true if researchers compute the survivor function, $S(t)$: they would need multiple covariate profiles and confidence intervals to assess x's effect.

overall process being modeled. That is, we may be interested not only in the risk of MID onset, but on the probability that two states will be involved in a MID at some point in time. Transition probabilities can be particularly useful in situations such as MIDs, because they aggregating a covariate's effect across the multiple transitions constituting the entire process. We discuss transition probabilities further in Appendix B.

Appendix G:
Model Performance Simulations: Details

A. Section II Simulations: Cox and L/P Performance, Basic Scenarios

1. SETUP DETAILS

We are interested in whether the Cox model can recover *x*'s true effect under a wide range of conditions. We perform the same checks for the incumbent correction in the literature: a logit model with cubic polynomials, as well as a logit model with natural cubic splines.[53] We run three broad classes of simulations that differ in the data's generating process (DGP). The DGPs have different baseline hazards, but otherwise, contain no assumption violations, i.e., no time-interacted covariates, or causal complexity, both of which we consider in more detail in the main text's Section IV.B. These are the simplest situations we could select, and in that way, our basic simulations favor logit.

The three classes of simulations we consider are:

1. Data in which the true DGP is a continuous-time duration setup. We use `survsim` (Crowther and Lambert 2012) to specify a number of different baseline hazard structures. We list and graph all the scenarios' baseline hazards later in this appendix, as well as in the simulation viewing app.[54]
2. All the baseline hazard scenarios from (1), but coerced into a de facto start-stop duration format.[55] This creates data in which the recorded durations are integer values. For instance, $t = 3.24$ would be forced to equal `ceil`(t) = 4. This class of simulations proxies how we usually work with duration data in political science: our process of interest is typically continuous in nature, but we *record* information about our process at more aggregated, discrete intervals (e.g., monthly or annually), yielding the canonical BTSCS data structure.

[53] We place three knots at *t*'s 10th, 50th, and 90th percentiles (Harrell 2015, 26–28), in addition to boundary knots at *t*'s minimum and maximum values. In situations where one of the boundary knots' values equaled an interior knot's value, we removed the relevant interior knot (see Appendix G, Sect. F for details on rates of occurrence).

[54] For logit, as we discuss later in Appendix G, Sect. C.1, we episode-split the continuous durations at observed failures, then use the split-*t* variable to generate the polynomials or splines.

[55] This format is synonymous with interval-censored duration data. We use "coerced start-stop" because the term is more descriptive for a wider audience.

3. Data in which the true DGP is discrete. We generate our scenarios using both the logit and complementary log-log (cloglog) link functions. Situations with a truly discrete DGP are rare in political science, but the Cox model assumes the underlying DGP is continuous, making this an important case to check. Notably, Cox models *can* handle truly discrete data by implementing the exact partial likelihood correction (`exactp`) for tied data (vs. the Breslow or Efron corrections) (Box-Steffensmeier and Jones 2004, 58).[56]

In all the simulations, *x* is time invariant within a subject's panel. For the scenarios in which we run logit on continuous-time data, we split the spells at each observed failure time in the dataset (i.e., `stsplit`). We then use the updated duration from these splits when we generate our time polynomials for each observation. We do this even though we never use logit in such applied situations.

We run each class of simulations for three different sample sizes: $N = \{100, 250, 500\}$. In total, there are 54 scenarios across all the different true DGP–baseline hazard form–sample size triplets: 21 from (1), 21 from (2), and 12 from (3). We provide the full set of simulation results in our supplemental materials. Here, we arbitrarily selected $N = 250$ to discuss, but the results are broadly similar across all the sample sizes.

2. Coefficient Estimates

[Insert Figure 10 here] – see end of this appendix

[Insert Figure 11 here] – see end of this appendix

Figures 10 and 11 display the results for the first and second classes of simulations, respectively. Each figure has three panels: a graph showing *x*'s estimated effect for each model compared to the true value (panel (a)), a graph showing the percent bias in each model's estimate of *x*'s effect (panel (b)), and a graph showing the root-mean-squared error (RMSE) for each model's *x* estimate (panel (c)). β_x's

[56] A Cox model with `exactp` for ties is equivalent to a conditional logit model with time dummies.

estimated confidence interval overlapping with β_x's true value (-1) indicates good performance (unbiased estimates) in panel (a). Shorter bars indicate better performance in panels (b) and (c).

Across both figures, the Cox model performs better than the logit when $N = 250$. When the true DGP is continuous, both the Cox's and all the logit's estimates are unbiased across all seven baseline hazard scenarios (Figure 10a). However, the Cox estimates' percent bias (Figure 10b) and RMSE (Figure 10c) are lower than those of both logit specifications, sometimes appreciably so (e.g., Scenario 4.1). As a semi-parametric model, the Cox's standard errors tend to be slightly larger than those from parametric models. The Cox's better RMSE performance is therefore telling. It means, despite the Cox's larger standard errors, its markedly better performance for percent bias makes it worth using over either logit specification.

When we coerce the data into start-stop format for all models, the results are similar (Figure 11).[57] All the models we estimate recover unbiased estimates across all seven baseline hazard scenarios. The Cox (Efron)'s performance continues to surpass that of either logit specification and of the Cox (`exactp`).[58] Again, the Cox's performance is telling, for the reasons we noted above *plus* the fact that coerced start-stop data represent the usual BTSCS data structure in political science.

On the whole, our simulations confirm that what we know in theory about the relationship between the Cox and logit bears out in practice. Both the Cox and logit recover unbiased estimates of x's true effect. As a slight wrinkle, we showed the Cox performs better across the board with less biased estimates, and that this improvement was of sufficient magnitude to offset the Cox's larger standard errors, producing the Cox's smaller RMSE. The simulations from the truly discrete DGPs tell the same

[57] We also estimate a Cox model with an exact partial tie correction (`exactp`) for the coerced start-stop data.

[58] More generally, researchers have investigated the Cox model's performance with interval-censored data. In some cases, Cox model estimates can be biased (e.g., Goggins et al. 1998; see Desmarais 2015, 7–9 for an overview). However, none of these studies assess whether logit estimates are *also* biased in these situations. We lay out some rough scope conditions under which the logit might outperform the Cox in Appendix I. Broadly speaking, such scenarios arise when the proportion of tied failure times is relatively high. We advocate more broadly for caution in these situations.

story, by and large (see simulation viewing app for graphs). The take-away from this set of simulations is simple: in situations with few ties, hedging your bets by estimating a Cox model instead of a logit has a near non-existent cost, across the continuous, start-stop, and discrete scenarios.

3. Predicted Quantities

Rainey's (2017) recent work points out methodologists should be just as concerned with an estimator's predicted quantities as with its actual coefficient estimates. He shows unbiased coefficient estimates are neither necessary nor sufficient to guarantee unbiased predicted quantities, especially in small samples. In the spirit of Rainey's larger point, we also run simulations in which we examine the Cox and logit's ability to recover accurate transition probabilities—our predicted quantity of choice from Cox models. We discuss these simulations' specifics in the next subsection (Appendix G, Sect. B), but they follow the same logic as Appendix G.A.2's simulations. There are 60 scenarios for each of our three classes of simulations.

The predicted quantity simulations return scant evidence challenging the Cox's performance. For most of the continuous-time scenarios, the Cox approximates the true transition probabilities. In situations where it does not (e.g., Scenario 5), the logit transition probabilities are generally even further from the truth. The Cox's good performance also holds for discrete time, where both Cox models and both logits return accurate transition probability estimates.

For the coerced start-stop scenarios, both Cox specifications and both logit specifications have more trouble recovering accurate transition probability estimates. Scenario 3.1 is particularly notable because it is one of the few scenarios where the two logits perform better than either Cox specification; the same performance patterns are present in Scenarios 4 and 4.1, but to a lesser extent. These scenarios are especially troublesome for the Cox model because they contain a particularly large number of tied failure times (i.e., subjects failing within the same $(t, t + 1]$ interval). While this is an important limitation of the Cox model, relative to L/P, such situations are likely to be fairly uncommon in most political science applications, and can be spotted prior to model selection.[59]

These scenarios aside, either all four models have no issue recovering accurate transition probability estimates *or* all four models have trouble. In short: our transition probability simulations also reaffirm our previous simulations: the estimators' theoretical properties, relative to one another, bears out in practice. Cox models typically perform as well, if not better than L/P, offering a strong rationale for researchers to employ them when modeling BTSCS data.

B. Transition Probabilities: Procedure

To simulate the transition probabilities from the Cox (Efron ties), Cox (exact partial ties), and logit models (splines, polynomials), we use the following procedure:

1. Define the DGP for the scenario.
2. Pull a dataset using (1)'s DGP.
3. Estimate the Cox (`efron`), Cox (`exactp`), logit (time polynomials), and logit (splines) models.
4. Compute each model's transition probabilities
    a. Cox: `mstsample` (Metzger and Jones 2018), one simulation draw only
    b. Logit: simulated computation based on modified code from `mstsample`
5. Repeat Steps 2–4 1000 times.

(For each model, there will now be 1000 columns filled with transition probability estimates for different time points.)

6. For each model's set of 1000 variables, average the 1000 variables' values for each time point (= point estimate)

[59] The Cox's subpar performance in all these scenarios makes sense, in light of our Appendix I discussions. For coerced start-stop, the Cox will typically be unbiased so long as less than ~50% of the subject-spells fail at any *t* value. As we also point out in Appendix I, these situations may arise on occasion, but are not frequently present in political science data.

7. For each model's set of 1000 variables, take the 2.5th and 97.5th percentiles for each time point (= conf. intvs.)

"x" symbols in the graphs denote where a point estimate's confidence interval does not overlap with the true transition probability. In some cases, when the estimated transition probabilities are near 1, the estimates show as being statistically different from the true transition probability. Technically, the truth and CIs indeed do not overlap, but the differences' magnitude are 0.001 or smaller.

This general procedure applies to all our simulation results (simple process, complex process). All these transition probability results are viewable in the supplemental material's viewing app.[60]

C. Setup Details: Section IV Simulations

1. Generalities

Each scenario is run with three different sample sizes: $N = \{100, 250, 500\}$. All simulations performed in either Stata 14.2/MP or 15.1/MP. We pull 1000 simulation draws unless noted otherwise. We use Crowther and Lambert's (2012) `survsim` package to generate the data, also unless noted otherwise.

For the coerced start-stop time scenarios, we round all the durations up to their nearest integer value. For instance, t = 3.24 would be forced to equal `ceil(`t`)` = 4. We then `stsplit` to create one observation for every integer time period, to mimic a BTSCS data structure.

Harden and Kropko (2019) detail a non-parametric approach for simulated duration data that generates its baseline hazard by fitting splines to randomly drawn points. As the authors admit, this situation deliberately favors the Cox, since the resultant baseline hazards can be very complex in form. Our use of simpler parametric functional forms for the baseline hazard, rather than Harden and Kropko's approach, further biases the simulations in favor of logit.

[60] App's location: https://bit.ly/3aFhGLd. Type `shiny::runApp()` once RStudio loads in the browser.

2. ADVANCED DGPS

For the Stage 1 → 2 transition, we use only a subset of our earlier baseline hazard scenarios, looking at only Scenarios 1 and 2 for coerced start-stop and Scenarios 7 and 9 for discrete. For Stage 2 → 1, we use either a monotonically increasing or monotonically decreasing hazard. We use a Weibull to generate these hazards for the coerced start-stop data and Scenarios 7 and 9 to do so for the discrete data. Forming the unique combinations of transition 1's $h_0(t)$ scenario and transition 2's produces 4 unique combinations.

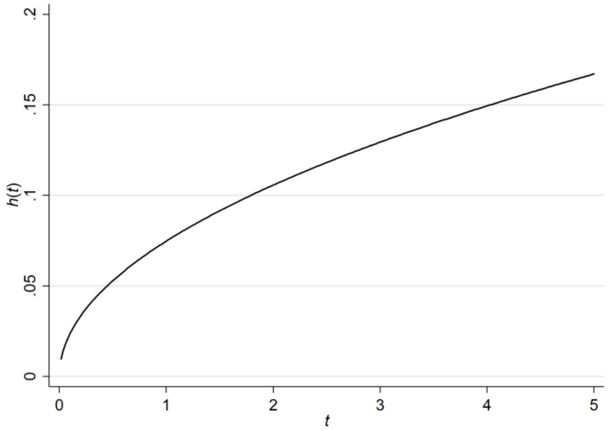
We run three different values of x's main effect {-0.2, 0, 0.2} and three different values of x's time-varying effect (TVE) {-0.2, 0, 0.2}. We do this for both transitions, meaning we have 4 * 3 (x_1 main) * 3 (x_1 TVE) * 3 (x_2 main) *3 (x_2 TVE) = 324 unique combinations per overarching DGP. With our two DGPs {coerced start-stop, discrete}, we have 648 different sets of simulation results. We use ln(t) as time's functional form for our TDEs.

Stage 1's natural cubic splines continue to have three knots at t's 10th, 50th, and 90th percentiles, plus the two boundary knots. However, we now compute t's percentiles using only stage 1 observations (hereafter notated $\text{perc}_{\text{stage\#}}$). We generated stage 2's natural cubic splines with one knot at t's $50^{\text{th}}_{\text{stage2}}$ percentile. We used fewer knots for stage 2 because of estimation issues for some of our scenarios. Few observations sometimes appeared in stage 2, which prevented the logit from converging when the specification contained many stage 2-related parameter estimates.

D. DGP: Continuous + Coerced Start-Stop

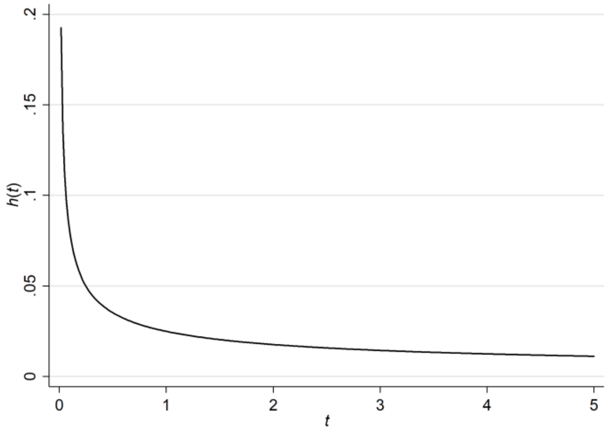
SCENARIO 1

- Description: Monotonically increasing hazard {Weibull}
- Command: `survsim t, l(1) g(1.5) covariates(x1 -1 cons -3)`
- *x*'s true effect: -1
- Baseline hazard graph:



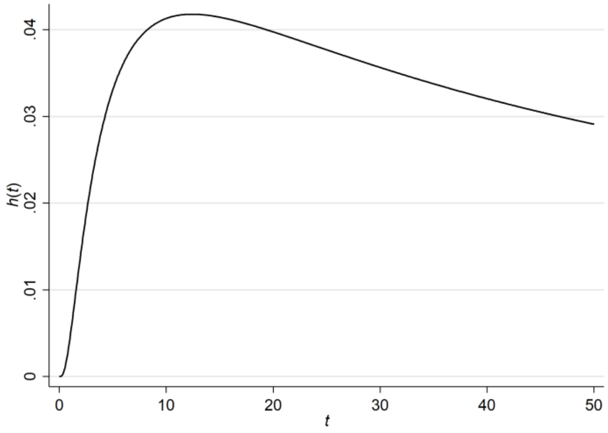
SCENARIO 2

- Description: Monotonically decreasing hazard {Weibull}
- Command: `survsim t, l(1) g(0.5) covariates(x1 -1 cons -3)`
- *x*'s true effect: -1
- Baseline hazard graph:



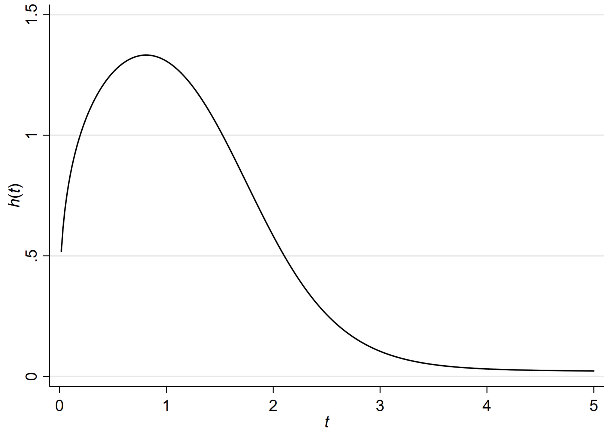
SCENARIO 3

- Description: Log-normal hazard
- Command: Manually generated; $t = \exp(3 - 1x + \sigma\varepsilon)$, where $\varepsilon \sim N(0,1)$
- x's true effect: -1 (in AFT, implying 1 for PH because $\sigma = 1$)
- Baseline hazard graph:



SCENARIO 3.1

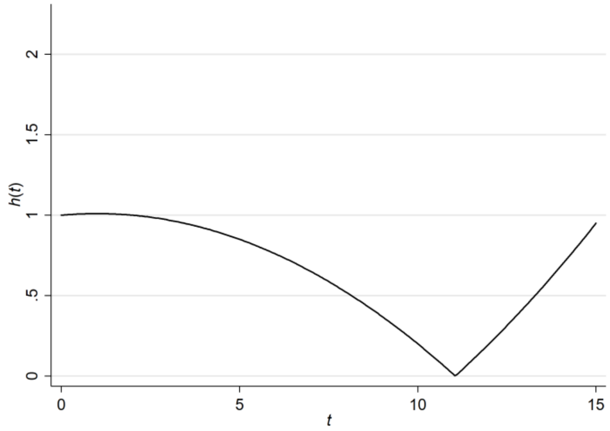
- Description: Concave down hazard {Weibull mixture}[61]
- Command: `survsim t, l(1.4 0.1) g(1.3 0.5) mixture pmix(0.9) maxtime(125) covariates(x1 -1 cons 0)`
- x's true effect: -1
- Baseline hazard graph:



[61] HT to Crowther and Lambert 2012, Fig. 1, bottom left panel.

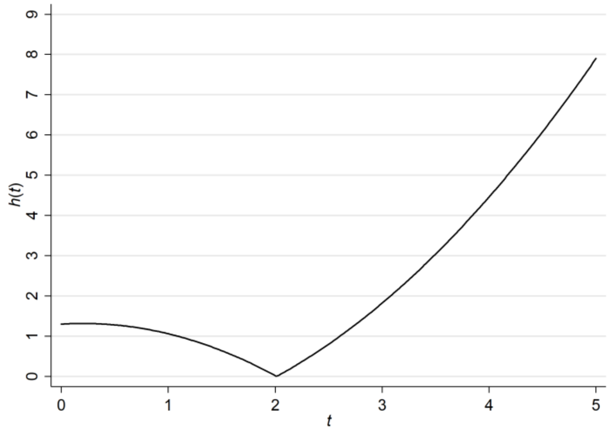
Scenario 4

- Description: Piecewise hazard #1 (long initial descent)
- Command:
  ```
  survsim t, hazard(abs(0.01:*#t:^2 :+ -0.02:*#t :+ -1))
  maxtime(35) covariates(x1 -1 cons 0)
  ```
- *x*'s true effect: -1
- Baseline hazard graph:



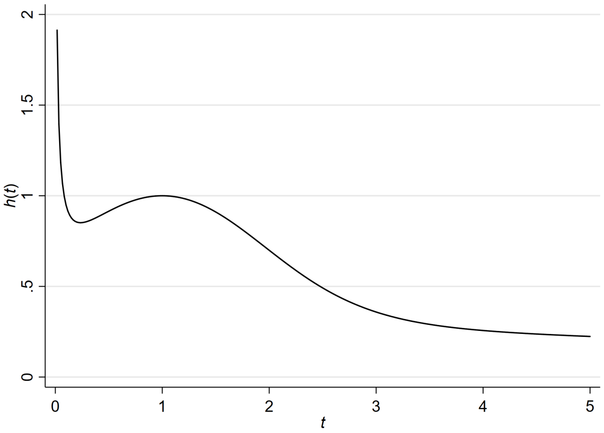
Scenario 4.1

- Description: Piecewise hazard #2 (quick initial descent)
- Command:
  ```
  survsim t, hazard(abs(0.4:*#t:^2 :+ -0.16:*#t :- 1.3))
  maxtime(35) covariates(x1 -1 cons 0)
  ```
- *x*'s true effect: -1
- Baseline hazard graph:



SCENARIO 5

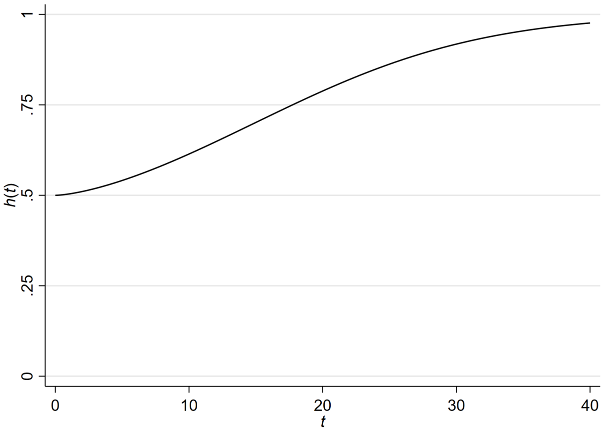
- Description: Complex hazard #1 {Weibull mixture}[62]
- Command: `survsim t, l(1 1) g(1.5 0.5) mixture pmix(0.5) maxtime(25) covariates(x1 -1 cons 0)`
- *x*'s true effect: -1
- Baseline hazard graph:



E. DGP: Discrete

SCENARIO 7

- Description: Monotonically increasing hazard, logit link[63]
- Command: Manually generated; $h(t) = \text{invlogit}(-3 - 1x + (0.06t)^{1.5})$
- *x*'s true effect: -1
- Baseline hazard graph:

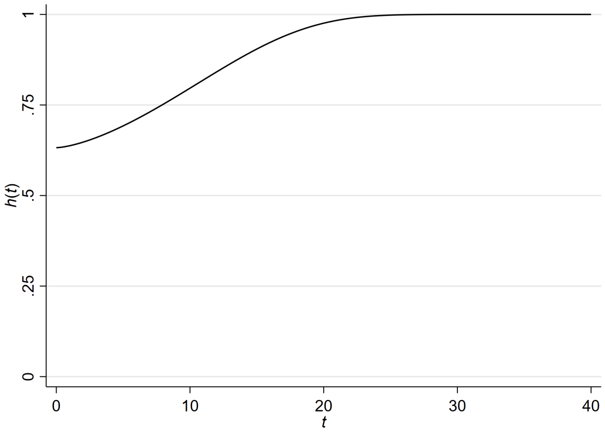


[62] HT to Crowther and Lambert 2012, Fig. 1, top left panel.

[63] Baseline hazard expression inspired by Carter and Signorino's replication code (2010).

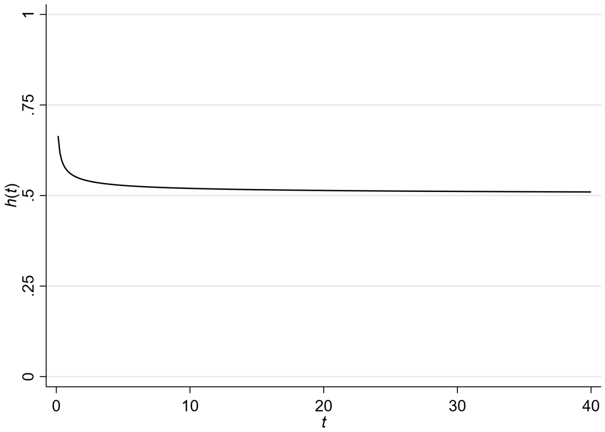
Scenario 8

- Description: Monotonically increasing hazard, complementary log-log link
- Command: Manually generated; $h(t) = \text{invcloglog}(-3 - 1x + (0.06t)^{1.5})$
- *x*'s true effect: -1
- Baseline hazard graph:



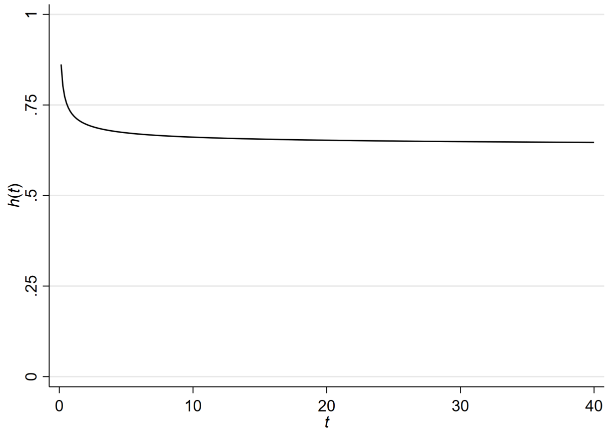
Scenario 9

- Description: Monotonically decreasing hazard, logit link
- Command: Manually generated; $h(t) = \text{invlogit}(-3 - 1x + 0.25t^{-0.5})$
- *x*'s true effect: -1
- Baseline hazard graph:



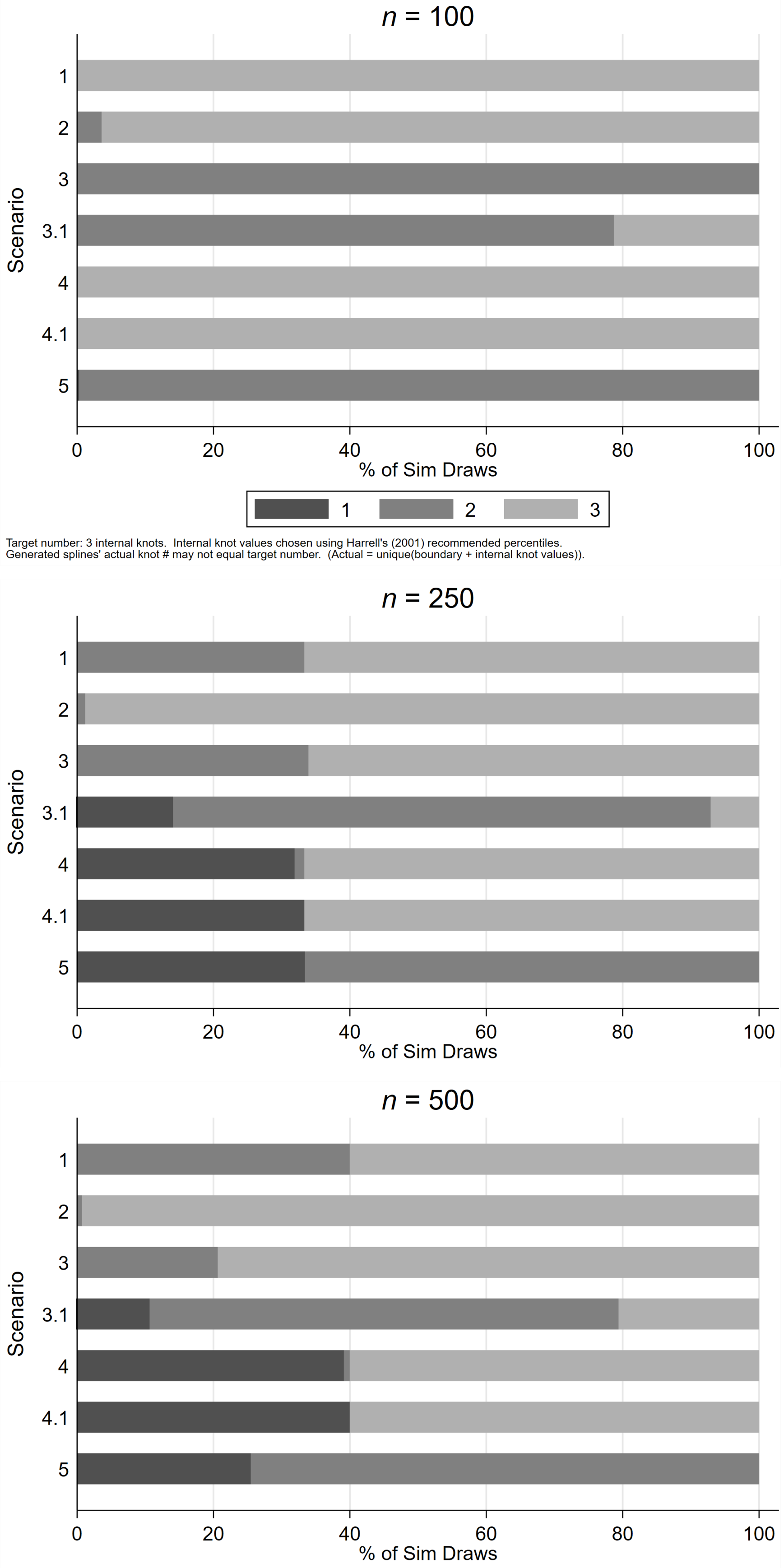
SCENARIO 10

- Description: Monotonically decreasing hazard, complementary log-log link
- Command: Manually generated; $h(t) = \text{invcloglog}(-3 - 1x + 0.25t^{-0.5})$
- *x*'s true effect: -1
- Baseline hazard graph:



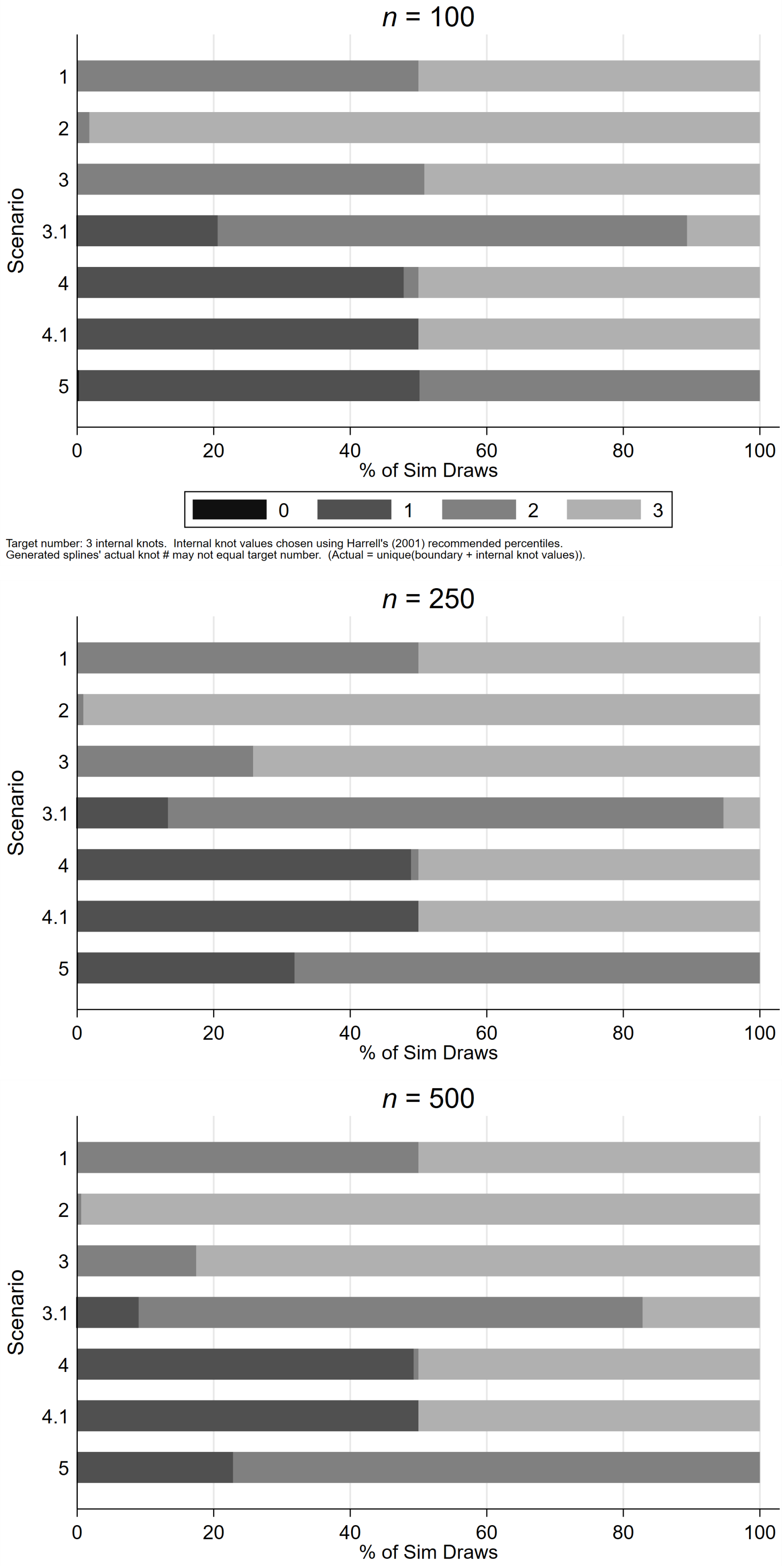
F. Number of Knots in Generated Natural Splines (Simple)

FIGURE 12. # of Actual Internal Knots for Logit Models with Time Splines, DGP = Continuous



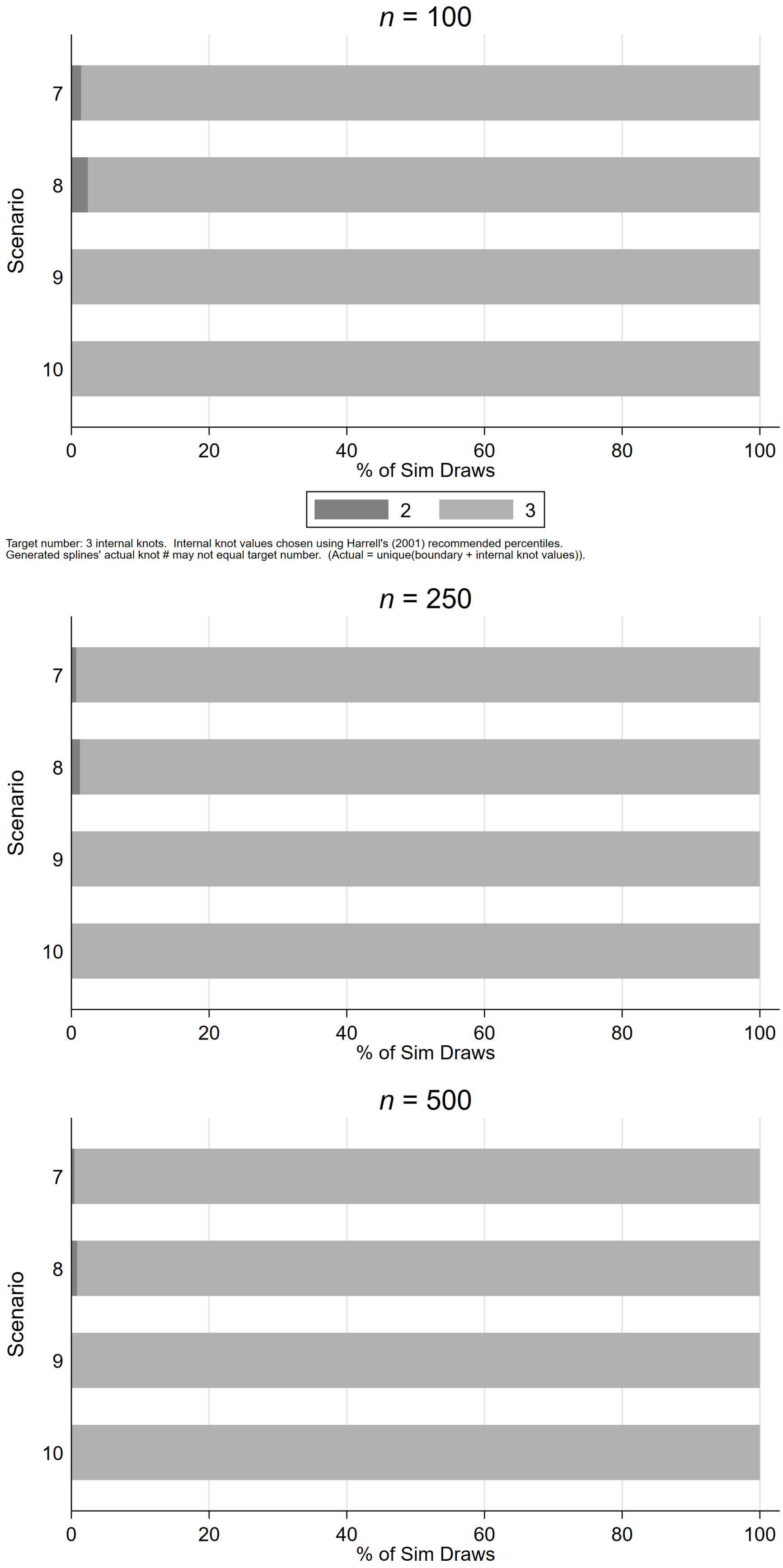
Target number: 3 internal knots. Internal knot values chosen using Harrell's (2001) recommended percentiles.
Generated splines' actual knot # may not equal target number. (Actual = unique(boundary + internal knot values)).

FIGURE 13. # of Actual Internal Knots for Logit Models with Time Splines, DGP = Coerced Start-Stop



Target number: 3 internal knots. Internal knot values chosen using Harrell's (2001) recommended percentiles.
Generated splines' actual knot # may not equal target number. (Actual = unique(boundary + internal knot values)).

FIGURE 14. # of Actual Internal Knots for Logit Models with Time Splines, DGP = Discrete

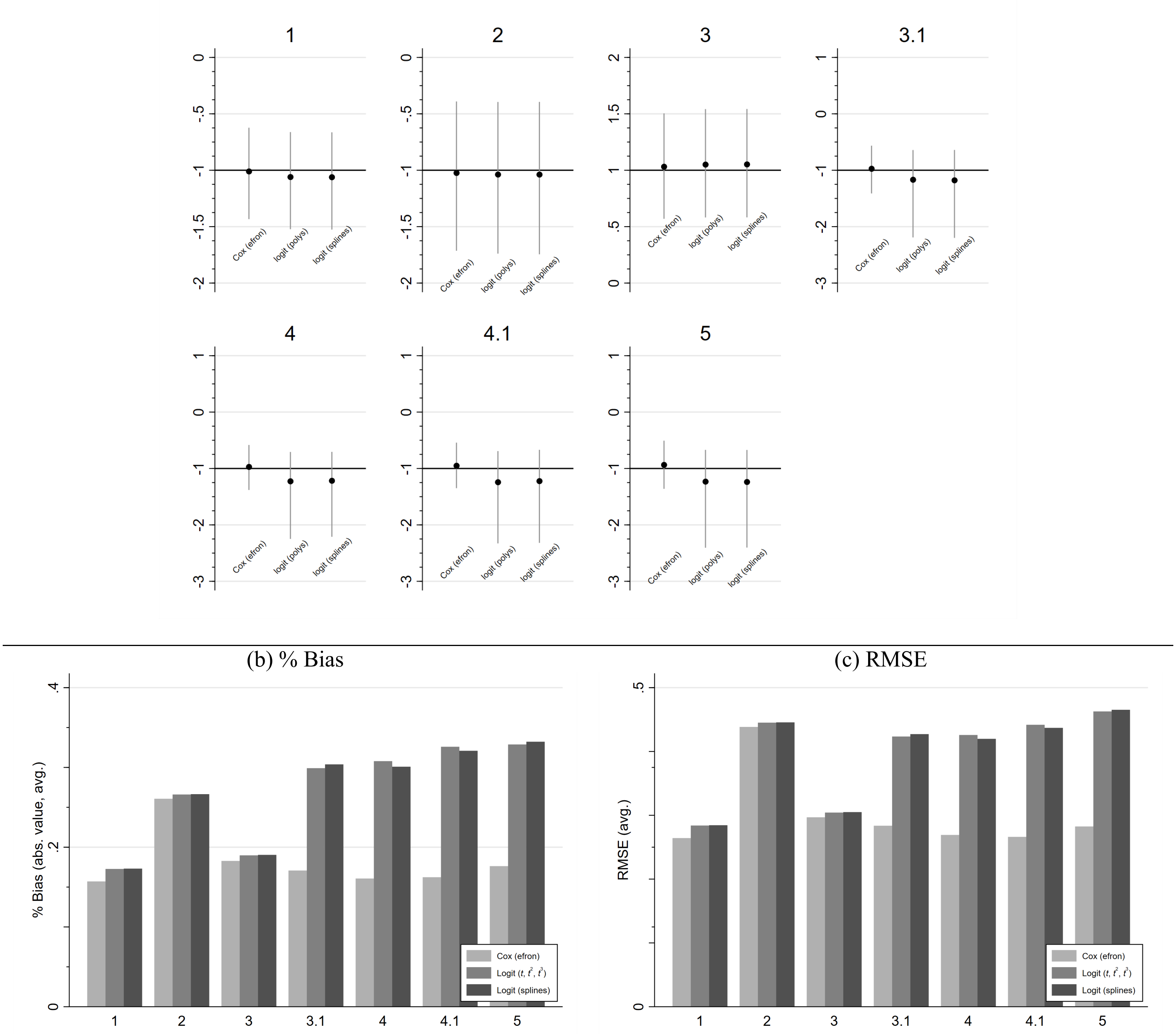


Target number: 3 internal knots. Internal knot values chosen using Harrell's (2001) recommended percentiles.
Generated splines' actual knot # may not equal target number. (Actual = unique(boundary + internal knot values)).

G. Number of Knots in Generated Natural Splines (Complex)

We deliberately opted not to generate similar graphs for our complex scenarios. Because of the unique combination of *x*'s main effect and time-varying effect, across two transitions, with two different baseline hazard patterns for transition 2, there would be 162 bars to graph per stage, yielding 324 bars per DGP class (vs. the simple scenario, where each DGP class had 12 bars only). Multiply that by two—one for coerced start-stop and one for discrete, and the end result is 648 bars to somehow arrange in a meaningful way. The cost in doing so outweighs the benefits, given the information's utility.

FIGURE 10. Simulation Results – Estimates, DGP = Continuous



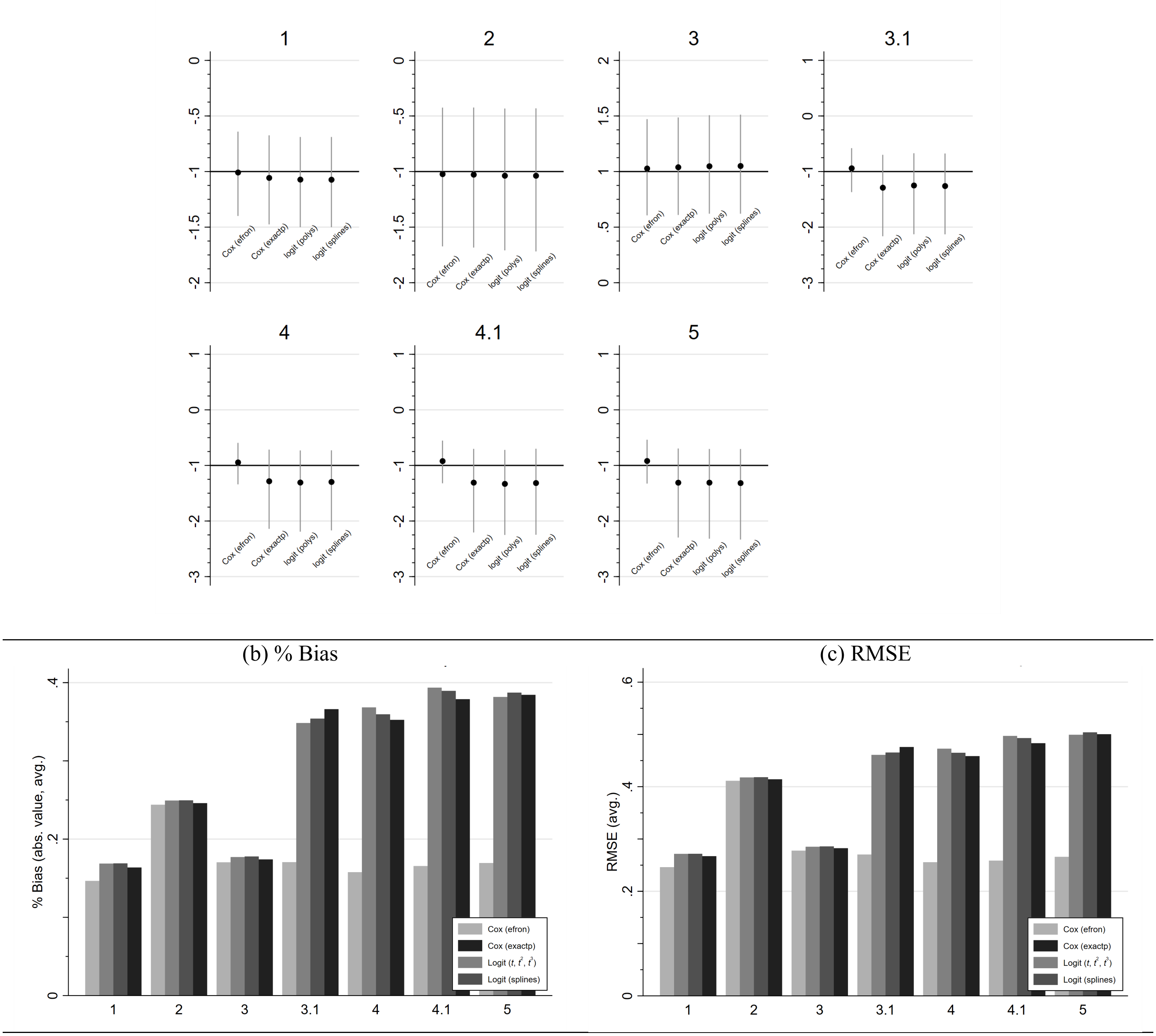
(a) Coefficient Estimates

(b) % Bias

(c) RMSE

N = 250. Quantities computed from 1000 simulations. 95% CIs reported in (a).

FIGURE 11. Simulation Results – Estimates, DGP = Coerced Start-Stop



(a) Coefficient Estimates

(b) % Bias

(c) RMSE

N = 250. Quantities computed from 1000 simulations. 95% CIs reported in (a).

Appendix H:
Parametric Duration Model Simulations

We primarily focus on the Cox model and how it can constitute an improvement over the typical way in which practitioners analyze BTSCS data in political science, logit or probit (L/P). Our results show using a Cox over a logit generally produces equivalent or better estimates of these effects, regardless of whether "effects" are conceptualized as coefficient estimates or predicted quantities. Using a Cox model also gives researchers more latitude to properly model more complex processes that might otherwise lead to erroneous conclusions. Further, because the Cox's estimates are generally unbiased, and have lower percent bias and a lower RMSE compared to logit, we are not simply advocating a "less bad" alternative to current practice, but an alternative that meets the criteria by which we judge an estimator's performance for a given DGP. However, in thinking about the Cox duration model in the context of BTSCS data, a natural follow-up is thinking about whether *other* types of duration models—namely, parametric duration models—might also provide similar improvements.

A. Parametric Model Overview

Parametric duration models differ from the Cox duration model in that they explicitly parameterize the baseline hazard, $h_0(t)$. In this way, parametric duration models and L/P are alike. The baseline hazard's functional form determines what kind of parametric duration model is used, with the exponential, Weibull, log-normal, log-logistic, Gompertz, and generalized gamma models being among the more common selections, but others exist.[64]

Unlike L/P, though, conventional parametric duration models' $h_0(t)$ functional forms are less intrinsically flexible in nature. L/P's time polynomials or cubic splines can approximate many different true baseline hazard forms, by virtue of all the potential functions one can express as a variant of a cubic time polynomial. By contrast, parametric duration models' $h_0(t)$s specify a general family of 'shapes,' and in fitting such a model, one forces the baseline hazard to follow that shape. For instance, the exponential duration model assumes a flat baseline hazard—whether an event occurs is completely

[64] See http://www.shawnakmetzger.com/parHaz for visuals of these six models' baseline hazards.

independent of how long a subject has been at risk.[65] Fitting an exponential duration model forces the estimated baseline hazard to be flat, regardless of whether or not it is in truth. As another example, a Weibull or Gompertz duration model permit a baseline hazard that either monotonically increases *or* monotonically decreases. The log-logistic and log-normal both assume a non-monotonic hazard that first increases, then decreases.[66]

In part because of the decreased flexibility of conventional parametric models, researchers have worked to formulate newer parametric model variants with baseline hazards that can encompass a wider range of situations. Royston and Parmar's (2002, henceforth "R&P") flexible parametric model is an important example. R&P's model uses natural cubic splines of ln(t) to model the baseline hazard and provides four different parameterization scales for covariate effects: proportional hazards, proportional odds (PO), probit, and Aranda-Ordaz's (1981) more generalized link function (of which PH and PO are special cases, corresponding to $\theta \rightarrow 0$ and = 1, respectively).

1. Potential Disadvantages

As we mention in the main text, social scientists are currently cautioned against using parametric duration models. Duration modelers argue the potential costs (misspecifying the baseline hazard, yielding inefficient or biased estimated quantities, as well as potentially biased predicted quantities) outweighs the possible benefits of doing so (general gains in efficiency). Several notable sources currently contend the Cox is preferable to use when substantive theory produces no strong predictions about the form of the baseline hazard[67,68] (e.g., Box-Steffensmeier and Jones 2004, 66–67; Cleves et al. 2010, 236; Golub

[65] An exponential parametric duration model is equivalent to a logit/probit model with no time-related covariates in its specification.

[66] The log-logistic is also capable of expressing monotonically decreasing hazards when its shape parameter (p) is ≤ 1.

[67] Some debate indirectly occurred about whether and when one should ascribe substantive import to the baseline hazard rate in political science in the late 1990s, spurred by Bennett (1999)'s response to Box-Steffensmeier and Jones (1997). The debate took place entirely on substantive grounds. Bennett's discussion is informational, with no applied example or simulations. BTSCS data do appear in Bennett's discussion, but only in the context of time-

2008a, 2008b; Mills 2011, 144–45; Singer and Willett 2003, 476; Ward and Ahlquist 2018, 230). Others hint at the Cox's preferability indirectly with statements about parametric models' suitability: e.g., "if the [baseline hazard] assumption is valid" or "the failure time distribution is assumed [to be] known" (Collett 2015, 148, 191; Harrell 2015, 475–76; Hosmer, Lemeshow, and May 2008, 244; Kalbfleisch and Prentice 2002, 95; Lawless 2002, 341).[69] As a corollary, the Cox's semi-parametric structure uses information only about the ordering of failure times when the model estimates, not the times' magnitude. As a result, the Cox is "less affected by outliers in the failures times than fully parametric methods" (Harrell 2015, 475).[70]

In addition to the potential costs of misspecifying $h_0(t)$, our own point in the main text about a model's capacity to address causal complexity also holds. We know the Cox generally performs better than logit for modeling causally complex processes because it estimates fewer parameters. The same logic would apply for parametric duration models. Each parametric duration model, aside from the exponential model, has at least one additional parameter to estimate as part of its baseline hazard specification. For each unique transition within a process, then, a parametric duration model would have at least one additional baseline hazard-related parameter.[71]

varying covariates. A similar debate about the ascribed meaning of the baseline hazard recurred in a BTSCS context in 2010 (Beck 2010; Carter and Signorino 2010).

[68] Parametric duration models appear more frequently in the physical sciences and engineering precisely because researchers' (mathematical) theories of the process being studied produce predictions about the underlying rate at which the event of interest occurs. Radioactive decay is an example.

[69] Some of these sources (e.g., Lawless) point out the importance of checking the Cox's PH assumption and frame the assumption as a major drawback of the Cox. However, the exponential, Weibull, and Gompertz parametric models also assume PH. The log-logistic model assumes proportional odds, similar to logit.

[70] Desmarais and Harden (2012) argue for using robust estimation techniques with the Cox to further mitigate the effect of outliers. They also provide a test to determine whether to estimate the Cox using conventional or robust methods.

[71] Flexible parametric models like Royston and Parmar's use splines to parameterize the baseline hazard. Everything we say about the logit with splines in Section IV.A of the main paper is therefore applicable to these models.

2. POTENTIAL ADVANTAGES

Nevertheless, the Cox does have drawbacks. First, we know that, if a researcher's goal is to produce a well-performing predictive model, parametric duration models are generally preferable because of their fully parametric functional form (see Appendix E). Second, we know that parametric duration models will be more precise than a Cox model if the duration's true $h_0(t)$ follows one of the canonical parametric distributions (Collett 2015, 148; Harrell 2015, 475). Third, we know the Cox's PH assumption is troublesome to some, and while several of the parametric models make a PH or PO assumption (fn. 69), not all do. The log-normal is the best known of the non-PH/non-PO conventional parametric duration models, but R&P's flexible parametric model parameterized using a probit link for the covariates is another example, as is their covariate parameterization using Aranda-Ordaz's (1981) more generalized link function for $\theta \in (0,1)$. From our work on this paper, we also know the Cox can perform poorly relative to logit in certain situations. Our Appendix I simulations reaffirm previous studies showing the Cox model's estimates are poor in the presence of many ties (Hertz-Picciotto and Rockhill 1997).

B. Simulation Setup

We run a small set of additional simulations that extend our simulations in fn. 18 of the main text, which entail a single transition (i.e. no causal complexity) and the presence of PH violations.[72] With them, we investigate whether parametric duration models outperform the Cox model in coerced start-stop situations, similar to most BTSCS encountered by political scientists. Additionally, the simulations allow us to check whether parametric models with no PH or PO assumption perform better than models that make these assumptions, in general. The simulations are identical to fn. 18 except for which models we run:

1. A probit model with time polynomials

[72] We run the same sets of parameter combinations as the main text, but we only report a set of representative graphs in text. The expanded graphs are not currently included in the supplemental viewing app.

2. A probit model with cubic splines
3. A Weibull duration model†
4. A flexible parametric model using a PH metric†
5. A flexible parametric model using a probit scale

†s denote a model that makes a PH assumption. No models in this list make a PO assumption. For both flexible parametric models, our natural cubic splines have 2 knots (vs. 3 knots for logit/probit). We use the same spline characteristics for the probit model as we do for our main simulations with the logits.

We estimate all five models with the appropriate time-varying effect correction (for doing so with the Weibull, see Zuehlke 2013).[73] This correction amounts to a PH correction for the two PH-assuming models (†s in above list). For the flexible parametric models, we interact *x* with all the spline terms when we attempt to model the time-varying effect properly. The flexible parametric model specifications did not converge for 6 of our 9 sets of parameter value sets in Scenario 2 because of the *x**2-knot spline interaction terms (e.g., Figure 15's bottom panel); we remove the flexible parametric estimates for all the Scenario 2 parameter sets for code generalizability reasons. The convergence trouble nicely illustrates one of our earlier points from the main text: models that parameterize the baseline hazard require more information to successfully return estimates than the semi-parametric Cox model. The current literature suggests the Weibull will perform the best for Scenarios 1 and 2, as these scenarios' true DGP is a Weibull with a PH violation.

For comparison, we also estimate a naïve log-normal parametric duration model. We convert our reported log-normal coefficients into a PH metric for comparability with the other coefficients. We do the same for probit by transforming our estimates by $\frac{\pi}{\sqrt{3}}$, putting them on the same scale as logit. For both the log-normal and probit transformation, we also convert the standard errors appropriately using the delta method. We run the same parameter value combinations as fn. 18's simulations in the main text, but we only report a set of representative graphs below.

[73] Our reported estimates cast a particularly optimistic light on the Weibull's performance because correcting for PH violations in a Weibull model is unusual in political science.

C. Simulation Results

[Insert Figure 15 about here]

[Insert Figure 16 about here]

1. Coefficient Estimates

Figure 15 displays the average estimate for *x*'s main effect from a set of representative simulations. The estimate's percent bias and RMSE appear in Figure 16. In general, we interpret these graphs with some caution. We know the probit, Weibull, and log-normal coefficients are expressed (or have been transformed to be expressed) on a comparable scale as the Cox and logit coefficients. It is less clear whether the flexible parametric model estimates also are on the same scale. For this reason, after we provide some basic interpretations for *x*'s main effect, we omit a discussion *x*'s time-varying effect and instead focus on the transition probabilities, as we know these quantities are on a comparable metric, regardless of model type.

We compare among the non-naïve models to make our judgments about performance. As the current literature suggests, the parametric models perform better than both Cox models for Scenarios 1 and 2 for recovering estimates of *x*'s main effect. Interestingly, instead of the Weibull, one of the two variants of the flexible parametric model always has the lowest percent bias and RMSE for Scenario 1 across all nine sets of parameter values we check. For Scenario 2, with the flexible parametric models out of the mix due to our aforementioned convergence issues, the Weibull fares somewhat better. It has the lowest percent bias in 8 of the 9 parameter sets, with the Cox having the lowest in the remaining set. However, for RMSE, the Cox has the smallest RMSE across all nine parameter sets. When we include the naïve log-normal results in our comparisons, it becomes the top performer for all of the Scenario 2 parameter sets, for both percent bias and RMSE. It also performs the best for both when no time-varying effect exists for Scenario 1.

2. Predicted Quantities

[Insert Figure 17 about here]

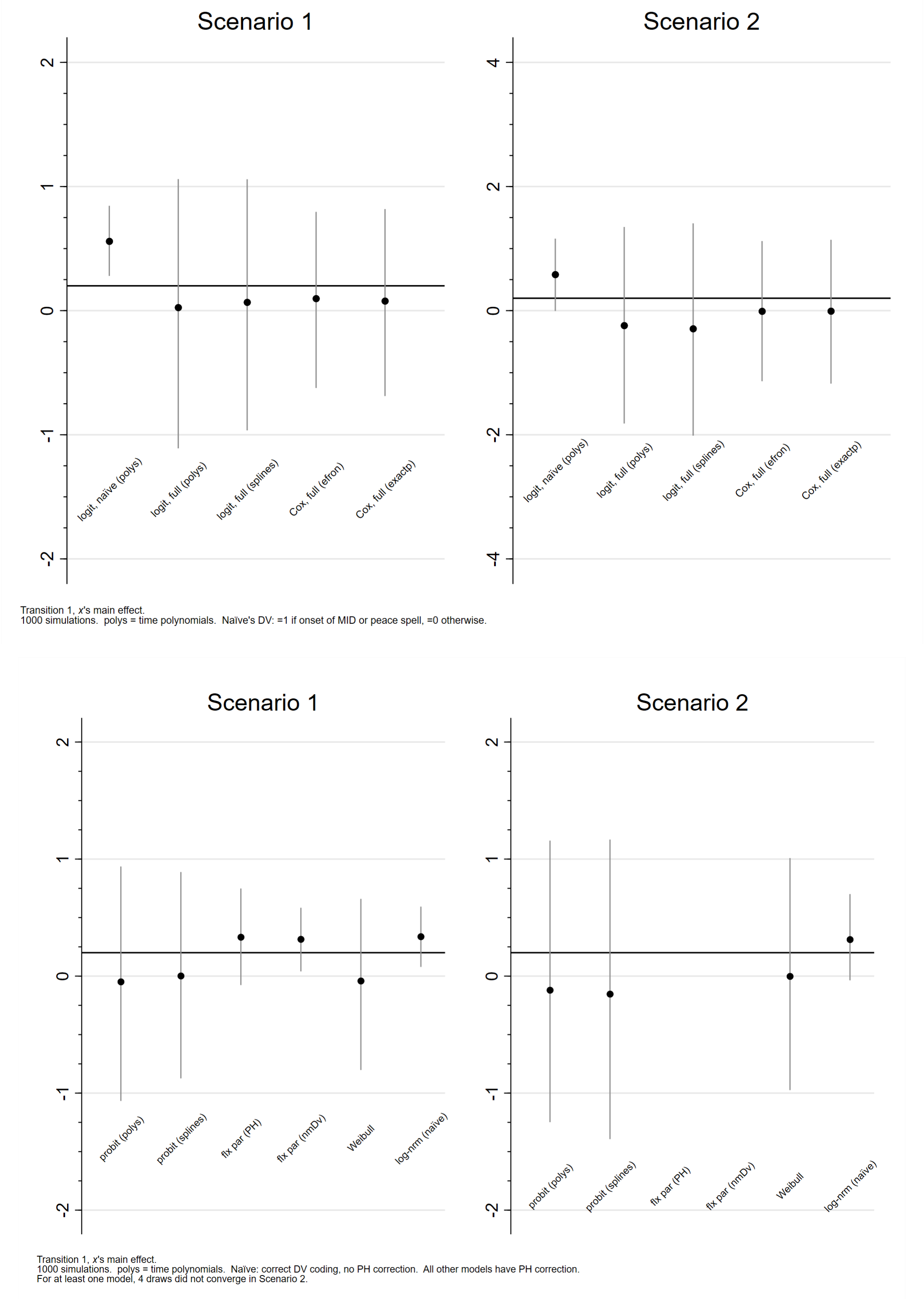
The transition probability results are more in line with our expectations from the main paper; the Cox performs even better than expected, in some cases. For Scenario 1, for instance, the Cox model (`efron`) is the only model whose transition probability CIs consistently include the true transition probability value for all 12 parameter value combinations, across the entire range $t \in (0,30]$ (Figure 17 for illustrative example). Among the models whose CIs include the true value for Scenario 1, the PH-corrected Weibull's estimate tends to be closest to the truth for $t \geq 20$, as the current literature would expect, though the true transition probability is near its upper bound for these t values (e.g., 0.99996). The Cox's behavior is more surprising in the current literature's eyes, as the simulations show its transition probability estimates are usually the closest to the truth in the middle t range. For smaller t values, the naïve log-normal's transition probability estimates are usually closest to the truth ($t \leq {\sim}6$, at minimum) (see logs included with supplemental materials).

[Insert Figure 18 about here]

For Scenario 2, the transition probabilities are unbiased for all 8 models we check, across all t values (Figure 18 for illustrative example). Unlike Scenario 1, Scenario 2 is in line with the current literature's expectations: one of the parametric duration model transition probability estimates is always closest to the true transition probability value for all of Scenario 2's parameter value combinations. The PH-corrected Weibull's transition probabilities tend to be closest to the true value for mid- to higher t values (usually $t \geq 16$, at least), with the naïve log-normal usually doing best for smaller *t*s (see supplemental material logs).

In summary, the general patterns we observe from the main text's simulations continue to be true across the two scenarios we check here: the Cox tends to perform just as well, if not better than, the other alternative modeling strategies we investigate. We place greater weight on our transition probability evidence because all these estimates are on a comparable metric, whereas it is less clear whether the same holds for the coefficient estimates. The Cox performs respectably regardless of scenario—its transition probability estimates are not always the closest to the true value, but across both scenarios, they are always unbiased for all $t \in (0,30]$ for all 12 parameter value sets we check. The same is not true of the parametric models. Their transition probability estimates outperform the Cox's in some situations (e.g., Figure 18), but are biased in others (e.g., Figure 17)—particularly notable because the true DGP for the two scenarios is a Weibull with a PH violation. On balance, then: the Cox is a safer, more versatile option, provided Appendix I's scope conditions are met, as they are for the two scenarios here.

FIGURE 15. Coerced Start-Stop, Coefficient Estimates, Overview: *x* = 0.2, *xln(*t*) = 0.2**

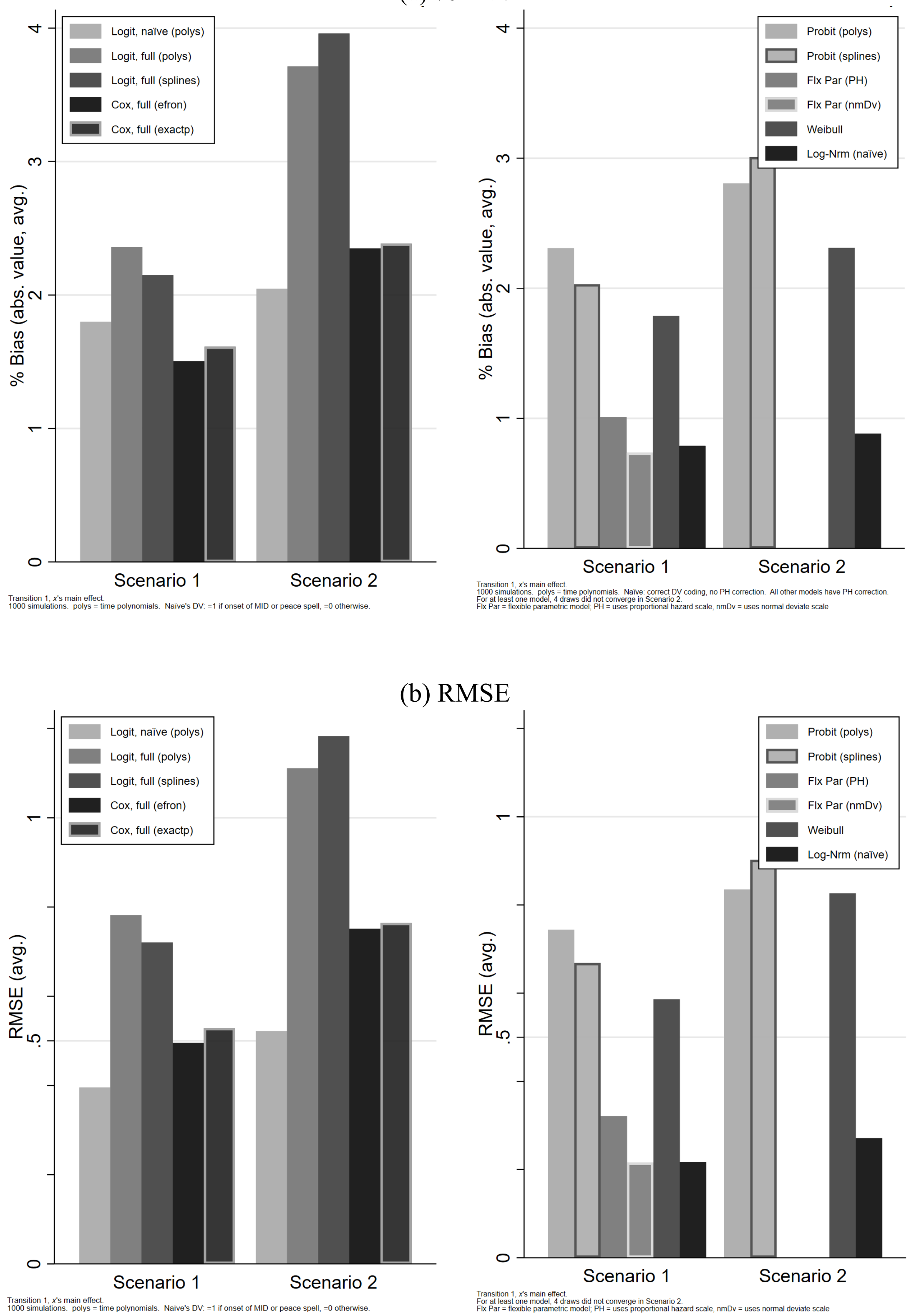


Transition 1, *x*'s main effect.
1000 simulations. polys = time polynomials. Naïve's DV: =1 if onset of MID or peace spell, =0 otherwise.

Transition 1, *x*'s main effect.
1000 simulations. polys = time polynomials. Naïve: correct DV coding, no PH correction. All other models have PH correction.
For at least one model, 4 draws did not converge in Scenario 2.

Notes: *N* = 250. Quantities computed from 1000 attempted simulations. 95% CIs reported. *y*-scales can differ across the 4 axes.

FIGURE 16. Coerced Start-Stop, Coefficient Estimates, % Bias and RMSE: $x = 0.2$, x*ln(t) = 0.2


(a) % Bias

(b) RMSE

Transition 1, *x*'s main effect.
1000 simulations. polys = time polynomials. Naïve's DV: =1 if onset of MID or peace spell, =0 otherwise.

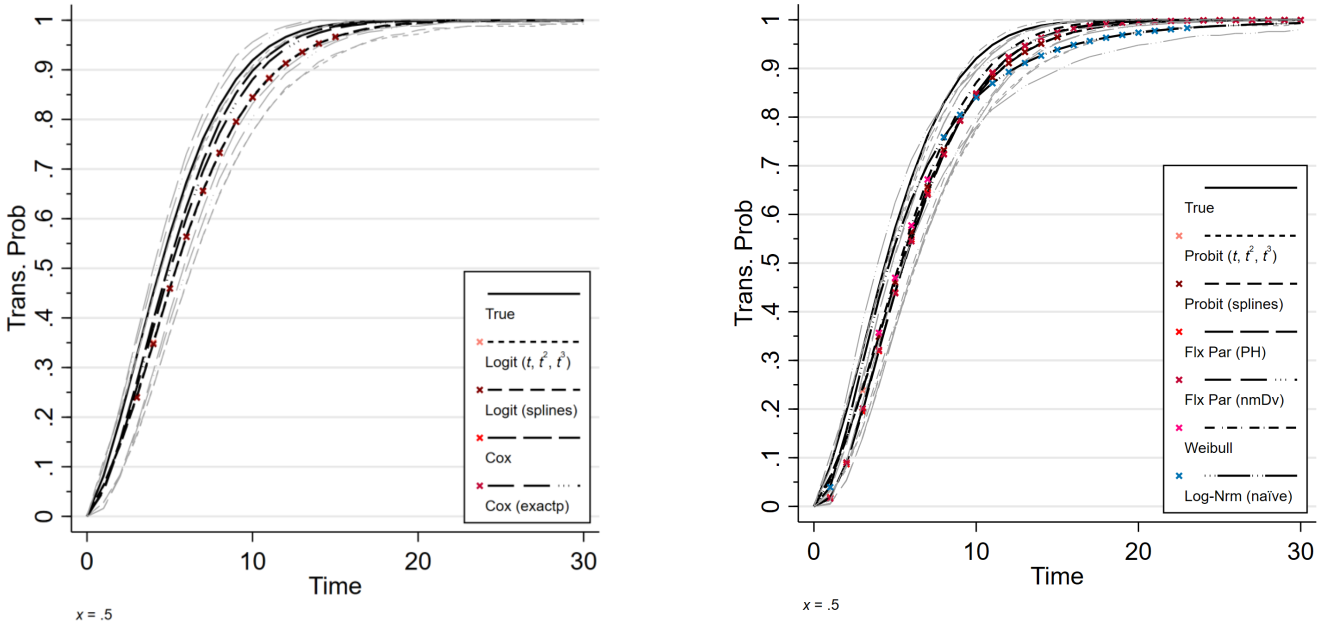
Transition 1, *x*'s main effect.
1000 simulations. polys = time polynomials. Naïve: correct DV coding, no PH correction. All other models have PH correction.
For at least one model, 4 draws did not converge in Scenario 2.
Flx Par = flexible parametric model, PH = uses proportional hazard scale, nmDv = uses normal deviate scale

Transition 1, *x*'s main effect.
1000 simulations. polys = time polynomials. Naïve's DV: =1 if onset of MID or peace spell, =0 otherwise.

Transition 1, *x*'s main effect.
For at least one model, 4 draws did not converge in Scenario 2.
Flx Par = flexible parametric model, PH = uses proportional hazard scale, nmDv = uses normal deviate scale

Notes: polys = polynomials. $N = 250$. Quantities computed from 1000 attempted simulations.

FIGURE 17. Coerced Start-Stop, Scenario 1: *x* = 0.2, *x**ln(*t*) = 0.2



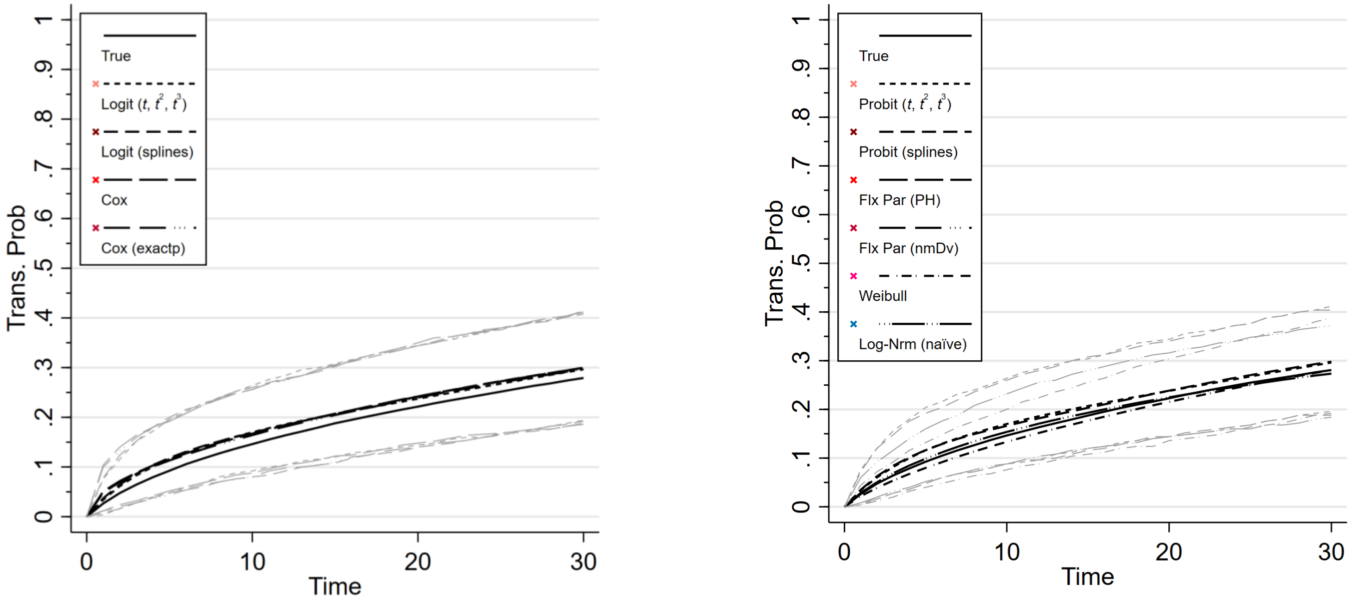
x = .5

Red Xs: Estimate stat diff from truth
nSubjs = 250, # unique failure times (avg.) = 23.115
# simulations attempted = 1000, # converged = 1000

x = .5

Red Xs: Estimate stat diff from truth
nSubjs = 250, # unique failure times (avg.) = 22.986
# simulations attempted = 1000, # converged = 1000
Flx Par = flexible parametric model; PH = uses proportional hazard scale, nmDv = uses normal deviate scale

FIGURE 18. Coerced Start-Stop, Scenario 2: *x* = 0.2, *x**ln(*t*) = 0.2



x = .5

Red Xs: Estimate stat diff from truth
nSubjs = 250, # unique failure times (avg.) = 28.041
# simulations attempted = 1000, # converged = 1000

x = .5

Red Xs: Estimate stat diff from truth
nSubjs = 250, # unique failure times (avg.) = 27.949
# simulations attempted = 1000, # converged = 996
Flx Par = flexible parametric model; PH = uses proportional hazard scale, nmDv = uses normal deviate scale

Appendix I:
Cox Model Performance: Scope Conditions

In the main text, as well as in Appendix G, we provide evidence that Cox models perform comparably to logit models under a number of scenarios for continuous time, coerced start-stop time, and discrete time. Of the scenarios we check, the Cox model's estimates practically always perform better than the logit's, and the same is generally true of the Cox's transition probabilities, compared to logit's.

However, our supportive evidence may be the byproduct of the specific DGPs we selected. The question becomes: under what conditions will the Cox return unbiased estimates for coerced start-stop durations, both for slope estimates and transition probabilities? To contextualize the discussion that follows, we provide a sense of what 'typical' political science data look like by calculating the mean/median durations, the percentage of subject-spells that fail at or before $t = 1$, and other related descriptive information for Appendix J's meta-analysis datasets (Table 9). We walk through the scope condition simulations in the next two sections before summarizing the practical implications in the final section.

A. Slope Estimates

Most prior studies of the Cox model's performance have focused on its ability to recover accurate covariate estimates. These studies have shown the Cox model's estimates are poor in the presence of many ties (Hertz-Picciotto and Rockhill 1997)—Cox model-speak for either "many subjects who experience the event at the same recorded failure time" or "few unique failure times at which the event occurs," depending on the source. In theory, if the DGP were truly continuous time and measured as such, ties would be non-existent as we could simply record increasingly precise failure times for each subject, but they nonetheless occur because of the frequency with which we gather our data (and the precision with which we can measure it, when we do). Kalbfleisch and Prentice go further, noting that "if…ties arise by the grouping of continuous failure times," the Cox's slope estimates will be inconsistent (2002, 106–7).

TABLE 9. Meta-Analysis: Data Characteristics

	Overview				*Non-Param. Estms.*			*Fails @ Modal* t *(Ties)*		
Article	*# Subj-Sp*	*Unique* t	*Total Fails*	*Max* t	*Mean*	*Median*	*10^{th} Perc.*	*# Fails*	*%* N_{fail}	*% Subj-Sp*
JOP2011_Beardsley	590	169	433	720	241.78	127	--	141	32.56%	23.90%
JOP2012_Aksoy	489	29	393	37	8.34	4	27	148	37.66%	30.27%
JOP2012_Kleinberg	14780	52	1724	53	46.72	--	--	632	36.66%	4.28%
JOP2013_Maves	296	38	179	63	22.62	13	--	19	10.61%	6.42%
JOP2014_Fuhrmann	1774	8	225	48	--*	--	--	214	95.11%	12.06%
JOP2014_Gibler	193	50	69	177	54.97	44	126	4	5.80%	2.07%
JOP2015_Hall	309	19	49	60	50.39	--	--	13	26.53%	4.21%
JOP2016_Bapat	216	22	137	37	18.06	13	--	39	28.47%	18.06%
JOP2016_Bayer	156	15	44	20	17.40	--	--	8	18.18%	5.13%
APSR2015_Berliner	64	32	32	2109	1144.28	1154	1997	1	3.13%	1.56%
APSR2016_Boushey	3730	47	1543	49	16.15	13	36	158	10.24%	4.24%
AJPS2010_Gilardi	1833	23	1823	23	2.00	2	4	115	6.31%	6.27%
AJPS2013_Way	190	15	18	55	49.41	--	--	3	16.67%	1.58%
AJPS2014_Fuhrmann	3209	44	239	51	46.62	--	--	38	15.90%	1.18%
AJPS2016_Arceneaux	102237	241	102234	717	427.86	459	675	1639	1.60%	1.60%
AJPS2016_Bapat	110	16	66	18	14.70	--	--	9	13.64%	8.18%
AJPS2016_Siroky	200	10	115	69	26.47	1	--	96	83.48%	48%

*: analysis uses weights; no restricted mean estimate

Overview

# Subj-Sp: count of unique subject-spells in dataset
Unique *t*: count of unique failure times in dataset
Total Fails (N_{fail}): count of observed failure times in entire estimation sample. If no ties, unique *t* = total fails.
Max *t*: largest recorded duration in dataset (regardless of failed or right censored)

Non-Parametric Estimates

Mean: mean duration (non-parametric, restricted mean)
Median: median duration (non-parametric); median is undefined if Kaplan-Meier estimate ($S(t)$) never drops below 50%.
10th Perc: 10th percentile (non-parametric), defined as $\min\{t \mid S(t) \leq 0.1\}$—10% of subjects will fail after this *t* value. Will be undefined if $S(\underline{t})$ never drops below 0.1.

Fails @ Modal

# Fails: count of observed failure events at the modal failure time
% N_{fail}: percentage of observed failure events falling at the modal failure time
% SS: percentage of subject-spells failing at the modal failure time

To reproduce this situation as our starting point, we vary the number of subjects failing within an interval by using what we know about $F(t)$. In datasets with no right censoring, $F(t)$ represents the percentage of failed subjects at the end of period t. If $F(t)$ is a straight, downward-sloping line, it means the same percentage of subjects fails on every unit interval—precisely what we need to produce a set number of failure events per unit interval. $F(t)$ becomes:

$$F(t) = \frac{(\%\ \text{fail})}{100} t$$

We use Bender, Augustin, and Blettner's (2005) inversion method to generate our data, exploiting the fact that *F*(*t*)'s range is [0,1] to solve for *t*:

$$t = \frac{100u}{\%\ \text{fail}}$$

where *u* ~uniform(0,1). We add a covariate to the DGP by recognizing that:

$$F(t) = F_0(t)^{\exp(X\beta)}$$

We treat our earlier *F*(*t*) expression as $F_0(t)$ and solve for *t*, yielding the final expression we use to generate our data:

$$\boxed{t = \frac{100}{\%\ \text{fail}}\left[1 - (1-u)^{1/\exp(X\beta)}\right]}$$

We include a single covariate, *x*, whose slope is equal to 1.25. As before, we coerce the resultant data into start-stop durations. We count from 5% of subjects failing in a unit interval up to 95%, in increments of 5%. We run 1000 simulations for sample sizes of 100, 250, and 500. Table 10 shows how many subjects fail for each sample size, in absolute terms. We report the output from three models as part of our supplemental app: (1) a Cox model using the true continuous-time duration as a baseline (`c` in the log output), (2) a Cox model with Efron ties using the coerced start-stop data (`ci` in output), and (3) a Cox model with exact partial ties using the coerced start-stop data (`cEi` in output).

TABLE 10. Number of Subjects Failing

	% Fail in Unit Interval									
n	5%	10%	15%	20%	25%	30%	35%	40%	45%	50%
100	5	10	15	20	25	30	35	40	45	50
250	12.5	25	37.5	50	62.5	75	87.5	100	112.5	125
500	25	50	75	100	125	150	175	200	225	250
S(*t*) = 0	20	10	7	5	4	4	3	3	3	2
	55%	60%	65%	70%	75%	80%	85%	90%	95%	
100	55	60	65	70	75	80	85	90	95	
250	137.5	150	162.5	175	187.5	200	212.5	225	237.5	
500	275	300	325	350	375	400	425	450	475	
S(*t*) = 0	2	2	2	2	2	2	2	2	2	

Our simulation results reaffirm what others have shown. First, the Cox performs well when the data are recorded as continuous time, regardless of how many fails occur on each unit interval. The Cox's good performance makes sense because truly continuous-time data have no tied failure times. Second, once we induce ties by coercing the times into a start-stop format, the Cox's coefficient estimates become increasingly biased as the number of ties increases, in line with previous research. The Cox using the Efron tie correction begins returning biased estimates with regularity when 50% or more of the subjects have the same failure time, across all three of our sample sizes.[74] With the exact partial tie correction, the Cox remains unbiased for longer. It never returns biased estimates for $N = 100$, but begins returning biased estimates when 85% of subjects have the same failure time(s) for $N = 250$ and when 70% of subjects have the same failure time(s) for $N = 500$. However, the fewer instances of bias are due to `exactp`'s far larger confidence intervals. In all 57 scenarios, its estimate's root-mean-squared error is larger than that of Cox (`efron`)'s, sometimes appreciably so. In 52 of the 57, `exactp`'s percent bias (the absolute value of it, specifically) is larger than `efron`'s. The five instances where `efron`'s |bias%| is larger, it is by small margins (0.005, 0.016, 0.027, 0.041, 0.094).

B. Transition Probabilities

We also have an interest in understanding when the model's predicted quantities will be biased. From the same simulations as above, we generate transition probabilities when $x = 0$. The same broad patterns emerge here as well, though in the case of Cox (`efron`), with slightly different scope conditions. First, like the slope estimates, the continuous-time Cox's estimated transition probability encompasses the true value for all integer t values in the output for all 57 scenarios.

Second, once we induce ties, the Cox (`efron`)'s transition probabilities become increasingly biased as the percentage of failures in a unit interval increases. Further, this bias appears *before* `efron`'s biased coefficient estimates do. The transition probabilities stop encompassing the true value for at least

[74] Because we have no right censoring, # fail/N_{total} = # fail/N_{fail} at any t.

one *t* value starting at around 30% of subjects have tied failure times for *N* = 100, 15% for *N* = 250, and 10% for *N* = 500 (vs. 50% for *β* across all three *N*s). For these specified ranges, the Cox (`efron`) consistently overestimates the true transition probability value.[75]

Notably, the biased *t* value is usually the last integer before *S*(*t*) = 0. For instance, when *N* = 100 and 30% of subjects fail in each interval (meaning *S*(4) = 0), *t* = 3 is biased for Cox (`efron`). In political science data, the *t* at which *S*(*t*) = 0 is often considerably larger than those from our simulation scenarios (Table 10's "S(*t*) = 0" row). Sometimes, *S*(*t*)'s value never reaches 0 in a dataset, if there are right-censored subjects whose recorded *t* is larger than the last recorded failure time. Table 9's "10th Perc." column displays the *t* value at which only 10% of subjects have a larger failure time. Eleven of the 17 replication articles have an undefined 10th percentile, meaning there is *never* a recorded *t* value after which 10% (or less) of subjects fail. The "Max *t*" column contains the largest recorded duration in each article's replication dataset, for additional context.

Third and finally, the Cox (`exactp`)'s transition probabilities have similar bias patterns as its coefficients once we induce ties. Its estimated transition probabilities' bias eventually becomes statistically different from zero, but not until the percentage of subjects failing at a given *t* is toward the upper end of the spectrum (90% for *N* = 100, 75% for *N* = 250, and 70% for *N* = 500).[76,77] The bias also continues to manifest far later than the Cox (`efron`)'s transition probability bias. However, also like Cox (`efron`), Cox (`exactp`)'s transition probability bias appears before its biased coefficients do, though the discrepancy between the two scenarios where the bias first manifests (in terms of percentage fails) is smaller for `exactp`. In general, `exactp` underestimates the transition probability's value (vs. `efron`, which overestimated it).

[75] More broadly, it overestimates the transition probability for all fail percentages and sample sizes, but the bias becomes statistically different from zero only for the specific ranges we mentioned.

[76] Unlike before, `exactp`'s transition probability CIs are the same general width as the Cox (`efron`)'s transition probability CIs, making quick comparisons to gauge relative performance easier.

[77] At all these percentages, there is only one failure time where the true transition probability is non-zero (*t* = 1), consistent with the biased transition probability estimate patterns for `efron`.

C. Practical Implications

Our scope condition simulations suggest two major takeaways for the start-stop Cox's performance: (1) issues generally begin to arise when 50% or more of subjects fail at the same t value, and (2) problems *can* arise with fewer than 50% failing at the same t if transition probabilities are also of interest.[78] The first echoes existing research on the Cox model's performance. In a political science context, though, a large number of ties on the order of 50% or more is relatively rare. Only two of Table 9's seventeen articles have more than 50% of its total failures occurring at a single time (% N_{fail} column).

Regarding our second takeaway, the transition probability evidence from these simulations also helps to better situate some of the main paper's simple process simulation results. There are some instances in which the Cox's transition probability estimates are biased for a coerced start-stop format, but logit's are not, like Scenarios 3.1, 4, and 4.1. With this appendix's simulation evidence in hand, the start-stop Cox's performance is now less surprising. Scenario 3.1's true baseline hazard rate is such that ~68% of subjects fail in $0 < t \leq 1$; Scenario 4's true $h_0(t)$, ~64% in $0 < t \leq 1$; and Scenario 4.1, ~65% in $0 < t \leq 1$ (see corresponding graphs in viewing app, solid line in $x = 0$ graph).

From these simulations alone, we are reluctant to conclude that many tied failure times is a sufficient condition for biased transition probabilities, and even more reluctant to associate a specific number with "many." Unlike the Cox's coefficient estimates, there is a paucity of simulation studies investigating the Cox's ability to recover accurate transition probability estimates under various conditions. Running such a set of simulations to somewhat exhaustively elucidate every condition is another paper in its own right. Putting those simulation results in context with the logit's performance under the same conditions is perhaps another. Thus, our goal in this appendix was to sketch out some

[78] Transition probabilities and traditional duration model quantities like $F(t)$ and $S(t)$ are related, with the last two being special cases of the first for simple Cox models. Therefore, the patterns we note for transition probability estimates likely apply to estimates of $F(t)$ and $S(t)$, too. The only potential source of difference is that we obtain our uncertainty estimates via simulation instead of analytically.

rough rules of thumb as a stopgap, based on a limited set of scenarios. We are more comfortable concluding that practitioners should more carefully weigh the merits of the Cox vs. logit when they have (a) few unique failure times and/or (b) a failure time or two with many observed events.

Appendix J:
Meta-Analysis of PH Tests

In this appendix, we perform a meta-analysis of all studies employing a logit, probit or rare events logit model with cubic polynomials for time in order to assess the prevalence of PH violations in existing studies. As described in the main text, we use Google Scholar to identify all research articles published in the *American Political Science Review*, the *American Journal of Political Science* and the *Journal of Politics* from 2010 to 2016 that cite Carter and Signorino (2010), and report at least one logit or probit model with time polynomials in the main text. Twenty-six articles match these parameters.[79] Of these 26 articles, replication materials are available for 17 articles.

Using these 17 articles as our sample, we begin by replicating the first logit or probit model with cubic polynomials reported in the text. In all instances, these models replicate without issue. Next, we re-estimate each of these 17 models as a Cox model. In each of the 17 cases, the Cox model returns very similar inferences to the logit or probit model, with the direction of all coefficients remaining constant, and only minor changes in *p*-values between the logit and Cox models.

Once the model is re-estimated as a Cox, we test for violations of the PH assumption using scaled Schoenfeld residual tests, which assess whether there are correlations between the residuals and some functional form of time (Therneau and Grambsch 2000). As Park and Hendry (2015) argue, which functional form of time researchers use is a potentially consequential choice that should be guided by an in-depth knowledge of the data and substance of a particular study. Specifically, when there are relatively few outliers in the data, Park and Hendry suggest the use of untransformed time, whereas the presence of outliers should lead researchers to use transformations of time (2015, 1085–86). While this type of in-depth knowledge of the data is valuable in applied research, it is beyond the scope of the present analysis. Instead, we perform the scaled Schoenfeld residual tests specifying three functional forms of time. First, we employ untransformed time (identity). Then, we employ two time transformations; the rank

[79] An additional five articles are similar in that they employ cubic time polynomials to account for duration dependence. However, these articles employ alternate statistical models including censored probits and multi-level models that cannot be directly replicated with a Cox model, and are therefore omitted from this analysis.

transformation, which Park and Hendry note is often the best choice of the transformed options (2015, 1086), and the log transformation, which is frequently employed in political science. By using three different functional forms of time, we are able to ensure that selecting one specification of time rather than another does not affect the PH test results. Importantly, regardless of which functional form of time is specified, the results of the PH tests are largely similar.

TABLE 11. Meta-Analysis of PH Violations

	Time Transformation		
	Identity	Rank	Log
Models with a global test violation	10 (58.8%)	10 (58.8%)	9 (52.9%)
Models with *at least* one coefficient violation	13 (76.5%)	12 (70.6%)	12 (70.6%)
Coefficients with a violation	52 (31.7%)	54 (32.9%)	54 (32.9%)

Percentages in parentheses are based on 17 total models, with 164 total coefficients across the models. All test results are based on $p \leq 0.1$ to indicate violations.

We present the results of the PH tests in Table 11. As the results indicate, regardless of which transformation of time is employed, the results of the PH tests are largely consistent. Across the 17 replicated models, 9 to 10 of the models return a statistically significant violation of the global test ($p \leq 0.1$), which is a test of the whether the combined effect of the covariates in the model violate the PH assumption (Therneau and Grambsch 2000). This is strong preliminary evidence that there may be violations of the PH assumption in roughly half of these models. However, as Box-Steffensmeier, Reiter and Zorn (2003) and Box-Steffensmeier and Jones (2004) note, in diagnosing PH violations, researchers should consider both the global test, and also individual covariate-specific effects. Both test results are important, as it may be the case that individual covariates violate the PH assumption, even if the global test fails to reject the null hypothesis. Thus, we also consider covariate-specific tests of nonproportionality. In so doing, we find that 12 to 13 of the 17 models have *at least* one covariate that violates the PH assumption ($p \leq 0.1$). Overall, of the 164 covariates across the 17 replicated models, we find that nearly a third of the covariates likely violate the PH assumption at the $p \leq 0.1$ level.

Some measure of caution is warranted in interpreting these results. Box-Steffensmeier and Jones (2004), Therneau and Grambsch (2000), and Park and Hendry (2015) all encourage researchers to use these diagnostic tests in conjunction with plots for each of the covariates to fully test for nonproportionality in covariate effects. However, the diagnostic tests for these replicated models provide strong suggestive evidence that nonproportionality is a prevalent feature in many political science studies, and a concern that merits rigorous and thorough testing in applied research.

Appendices Works Cited

Aalen, Odd, Ornulf Borgan, and Hakon Gjessing. 2008. *Survival and Event History Analysis: A Process Point of View*. New York: Springer.

Aberson, Christopher L. 2019. *Applied Power Analysis for the Behavioral Sciences*. 2nd ed. New York: Routledge.

Andersen, Per Kragh, Ornulf Borgan, Richard D. Gill, and Niels Keiding. 1993. *Statistical Models Based on Counting Processes*. New York: Springer.

Aranda-Ordaz, Francisco J. 1981. "On Two Families of Transformations to Additivity for Binary Response Data." *Biometrika* 68 (2): 357–63.

Austin, Peter C. 2018. "Statistical Power to Detect Violation of the Proportional Hazards Assumption When Using the Cox Regression Model." *Journal of Statistical Computation and Simulation* 88 (3): 533–52.

Austin, Peter C., and Janet E. Hux. 2002. "A Brief Note on Overlapping Confidence Intervals." *Journal of Vascular Surgery* 36 (1): 194–95.

Beck, Nathaniel. 2010. "Time Is Not A Theoretical Variable." *Political Analysis* 18 (3): 293–94.

Beck, Nathaniel, Jonathan Katz, and Richard Tucker. 1998. "Taking Time Seriously: Time-Series-Cross-Section Analysis with a Binary Dependent Variable." *American Journal of Political Science* 42 (4): 1260–88.

Bender, Ralf, Thomas Augustin, and Maria Blettner. 2005. "Generating Survival Times to Simulate Cox Proportional Hazards Models." *Statistics in Medicine* 24 (11): 1713–23.

Bennett, D. Scott. 1999. "Parametric Models, Duration Dependence, and Time-Varying Data Revisited." *American Journal of Political Science* 43 (1): 256–70.

Beyersmann, Jan, Arthur Allignol, and Martin Schumacher. 2011. *Competing Risks and Multistate Models with R*. New York: Springer.

Box-Steffensmeier, Janet M., and Bradford S. Jones. 1997. "Time Is of the Essence: Event History Models in Political Science." *American Journal of Political Science* 41 (4): 1414–61.

------. 2004. *Event History Modeling: A Guide for Social Scientists*. Cambridge: Cambridge University Press.

Box-Steffensmeier, Janet M., Dan Reiter, and Christopher J.W. Zorn. 2003. "Nonproportional Hazards and Event History Analysis in International Relations." *Journal of Conflict Resolution* 47 (1): 33–53.

Box-Steffensmeier, Janet M., and Christopher J. W. Zorn. 2002. "Duration Models for Repeated Events." *Journal of Politics* 64 (4): 1069–94.

Cameron, A. Colin, and Pravin K. Trivedi. 2005. *Microeconometrics: Methods and Applications*. Cambridge University Press.

Carter, David B., and Curtis S. Signorino. 2010. "Back to the Future: Modeling Time Dependence in Binary Data." *Political Analysis* 18 (3): 271–92.

Cleves, Mario, William Gould, Roberto Gutierrez, and Yulia Marchenko. 2010. *An Introduction to Survival Analysis Using Stata*. 3rd ed. College Station, TX: Stata Press.

Collett, David. 2015. *Modelling Survival Data in Medical Research*. 3rd ed. Boca Raton, FL: Chapman and Hall/CRC.

Cox, David R. 1972. "Regression Models and Life-Tables." *Journal of the Royal Statistical Society, B* 34 (2): 187–220.

Crowther, Michael J., and Paul C. Lambert. 2012. "Simulating Complex Survival Data." *Stata Journal* 12 (4): 674–87.

Dabrowska, Dorota. 1995. "Estimation of Transition Probabilities and Bootstrap in a Semiparametric Markov Renewal Model." *Journal of Nonparametric Statistics* 5 (3): 237–59.

Desmarais, Bruce A. 2015. "Discrete Measurement of Time and Interval Censoring in Event History Analysis." Working paper. http://papers.ssrn.com/abstract=2614922.

Desmarais, Bruce A., and Jeffrey J. Harden. 2012. "Comparing Partial Likelihood and Robust Estimation Methods for the Cox Regression Model." *Political Analysis* 20 (1): 113–35.

Gill, Richard D., and Soren Johansen. 1990. "A Survey of Product-Integration with a View Toward Application in Survival Analysis." *The Annals of Statistics* 18 (4): 1501–55.

Goggins, William B., Dianne M. Finkelstein, David A. Schoenfeld, and Alan M. Zaslavsky. 1998. "A Markov Chain Monte Carlo EM Algorithm for Analyzing Interval-Censored Data under the Cox Proportional Hazards Model." *Biometrics* 54 (4): 1498–1507.

Golub, Jonathan. 2008a. "Survival Models." In *Oxford Handbook of Political Methodology*, eds. Janet M. Box-Steffensmeier, Henry E. Brady, and David Collier. Oxford: Oxford University Press, pp. 530–46.

------. 2008b. "The Study of Decision-Making Speed in the European Union: Methods, Data and Theory." *European Union Politics* 9 (1): 167–79.

Hall, Matthew E. K., and Joseph Daniel Ura. 2015. "Judicial Majoritarianism." *Journal of Politics* 77 (3): 818–32.

Harden, Jeffrey J., and Jonathan Kropko. 2019. "Simulating Duration Data for the Cox Model." *Political Science Research and Methods* 7 (4): 921–28.

Harrell, Frank E. 2015. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. 2nd ed. New York: Springer.

Hertz-Picciotto, Irva, and Beverly Rockhill. 1997. "Validity and Efficiency of Approximation Methods for Tied Survival Times in Cox Regression." *Biometrics* 53 (3): 1151–56.

Hosmer, David W., Stanley Lemeshow, and Susanne May. 2008. *Applied Survival Analysis: Regression Modeling of Time to Event Data*. 2nd ed. Hoboken, NJ: Wiley.

Hougaard, Philip. 2000. *Analysis of Multivariate Survival Data*. New York: Springer.

Jin, Shuai, and Frederick J. Boehmke. 2017. "Proper Specification of Nonproportional Hazards Corrections in Duration Models." *Political Analysis* 25 (1): 138–44.

Jones, Benjamin T., and Shawna K. Metzger. 2019. "Different Words, Same Song: Advice for Substantively Interpreting Duration Models." *PS: Political Science & Politics* 52 (4): 691–95.

Kalbfleisch, John D., and Ross L. Prentice. 2002. *The Statistical Analysis of Failure Time Data*. 2nd ed. Hoboken, NJ: Wiley-Interscience.

Lawless, Jerald F. 2002. *Statistical Models and Methods for Lifetime Data*. 2nd ed. Hoboken, NJ: Wiley-Interscience.

Long, J. Scott. 1997. *Regression Models for Categorical and Limited Dependent Variables*. Thousand Oaks, CA: Sage.

Metzger, Shawna K., and Benjamin T. Jones. 2016. "Surviving Phases: Introducing Multistate Survival Models." *Political Analysis* 24 (4): 457–77.

------. 2018. "`mstatecox`: A Package for Simulating Transition Probabilities from Semiparametric Multistate Survival Models." *Stata Journal* 18 (3): 533–63.

Mills, Melinda. 2011. *Introducing Survival and Event History Analysis*. Los Angeles: Sage.

Park, Sunhee, and David J. Hendry. 2015. "Reassessing Schoenfeld Residual Tests of Proportional Hazards in Political Science Event History Analyses." *American Journal of Political Science* 59 (4): 1072–87.

Rainey, Carlisle. 2017. "Transformation-Induced Bias: Unbiased Coefficients Do Not Imply Unbiased Quantities of Interest." *Political Analysis* 25 (3): 402–9.

Royston, Patrick, and Mahesh K. B. Parmar. 2002. "Flexible Parametric Proportional-Hazards and Proportional-Odds Models for Censored Survival Data, with Application to Prognostic Modelling and Estimation of Treatment Effects." *Statistics in Medicine* 21 (15): 2175–97.

Singer, Judith D., and John B. Willett. 2003. *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*. Oxford: Oxford University Press.

Therneau, Terry M., and Patricia M. Grambsch. 2000. *Modeling Survival Data: Extending the Cox Model*. New York: Springer.

Ward, Michael D., and John S. Ahlquist. 2018. *Maximum Likelihood for Social Science: Strategies for Analysis*. Cambridge: Cambridge University Press.

Wreede, Liesbeth C. de, Marta Fiocco, and Hein Putter. 2010. "The Mstate Package for Estimation and Prediction in Non- and Semi-Parametric Multi-State and Competing Risks Models." *Computer Methods and Programs in Biomedicine* 99 (3): 261–74.

------. 2011. "mstate: An R Package for the Analysis of Competing Risks and Multi-State Models." *Journal of Statistical Software* 38 (7): 1–30.

Zuehlke, Thomas W. 2013. "Estimation and Testing of Nonproportional Weibull Hazard Models." *Applied Economics* 45 (15): 2059–66.