Supplemental material for
“Implementation Matters: Evaluating the Proportional Hazard Test’s Performance”
Political Analysis

Shawna K. Metzger
University at Buffalo
smetzger@buffalo.edu

Appendix A
$\mathcal{J}^{\mathrm{E}}$ for J = 2: Approximation vs. Actual

To start, remember the general form of the expanded information matrix ($\mathcal{J}^{\mathrm{E}}$) (main text, Sect. II.A):

$$\mathcal{J}^{\mathrm{E}} = \begin{pmatrix} \mathcal{J}_1 & \mathcal{J}_2 \\ \mathcal{J}_2' & \mathcal{J}_3 \end{pmatrix} \qquad \mathcal{J}_1 = \sum \hat{V}(t_k)$$

$$\mathcal{J}_2 = \sum \hat{V}(t_k) g(t_k)$$

$$\mathcal{J}_3 = \sum \hat{V}(t_k) g^2(t_k)$$

In a Cox model with two covariates, after the $\bar{V}$ substitution but before demeaning $g(t_k)$, the approximated PH test's $\mathcal{J}^{\mathrm{E}}$ will equal:

$$\mathcal{J}^{\mathrm{E}^{apx}} = \begin{bmatrix} \frac{\sum_{k=1}^{K} \hat{V}(t_k, x_1)}{d} & \frac{\sum_{k=1}^{K} \widehat{\mathrm{Cov}}(t_k)}{d} & 0 & 0 \\ \frac{\sum_{k=1}^{K} \widehat{\mathrm{Cov}}(t_k)}{d} & \frac{\sum_{k=1}^{K} \hat{V}(t_k, x_2)}{d} & 0 & 0 \\ 0 & 0 & \frac{\sum_{k=1}^{K} \hat{V}(t_k, x_1) \sum_{k=1}^{K} g^2(t_k)}{d} & \frac{\sum_{k=1}^{K} \widehat{\mathrm{Cov}}(t_k) \sum_{k=1}^{K} g^2(t_k)}{d} \\ 0 & 0 & \frac{\sum_{k=1}^{K} \widehat{\mathrm{Cov}}(t_k) \sum_{k=1}^{K} g^2(t_k)}{d} & \frac{\sum_{k=1}^{K} \hat{V}(t_k, x_2) \sum_{k=1}^{K} g^2(t_k)}{d} \end{bmatrix}$$

where $\widehat{\mathrm{Cov}}(t_k) \equiv \widehat{\mathrm{Cov}}(t_k, x_1, x_2)$. Recall that $\bar{V} = \sum_{k=1}^{K} \hat{V}(t_k) \,/d$ (or, for the off-diagonals within a block, $\sum_{k=1}^{K} \widehat{\mathrm{Cov}}(t_k) \,/d$). This means $\bar{V}$ is constant across all $k \in K$ failure times. As a result, we can bring it outside the $g^2(t_k)$ summation in $\mathcal{J}_3$ (Therneau 2021, lines 38–41).

By contrast, the actual PH test calculation's $\mathcal{J}^{\mathrm{E}}$ for the same model will equal the matrix on the next page. For $\mathcal{J}_3$, unlike the approximation, we cannot factor out $\hat{V}(t_k)$ or $\widehat{\mathrm{Cov}}(t_k)$ from the $g^2(t_k)$ summation. Those quantities' values are now potentially different at each k, as we are no longer computing their average across all the observed failure times. This inability to factor is why $g(t)$ only appears *inside* the k-summations in the main text's Eq. (4).

The matrix below—specifically, its inverse—is only one part of the final PH test calculation. I walk through the full computation in Appendix B.

$$\mathcal{J}^{\mathrm{E}^{act}}=\begin{bmatrix} \sum_{k=1}^{K}\hat{V}(t_k,x_1) & \sum_{k=1}^{K}\widehat{\mathrm{Cov}}(t_k) & \sum_{k=1}^{K}\hat{V}(t_k,x_1)g(t_k) & \sum_{k=1}^{K}\widehat{\mathrm{Cov}}(t_k)g(t_k) \\ \sum_{k=1}^{K}\widehat{\mathrm{Cov}}(t_k) & \sum_{k=1}^{K}\hat{V}(t_k,x_2) & \sum_{k=1}^{K}\widehat{\mathrm{Cov}}(t_k)g(t_k) & \sum_{k=1}^{K}\hat{V}(t_k,x_2)g(t_k) \\ \sum_{k=1}^{K}\hat{V}(t_k,x_1)g(t_k) & \sum_{k=1}^{K}\widehat{\mathrm{Cov}}(t_k)g(t_k) & \sum_{k=1}^{K}\hat{V}(t_k,x_1)g^2(t_k) & \sum_{k=1}^{K}\widehat{\mathrm{Cov}}(t_k)g^2(t_k) \\ \sum_{k=1}^{K}\widehat{\mathrm{Cov}}(t_k)g(t_k) & \sum_{k=1}^{K}\hat{V}(t_k,x_2)g(t_k) & \sum_{k=1}^{K}\widehat{\mathrm{Cov}}(t_k)g^2(t_k) & \sum_{k=1}^{K}\hat{V}(t_k,x_2)g^2(t_k) \end{bmatrix}$$

Appendix B
Calculating T_j^{act}

Metzger (2023c, Appx. A) works through the approximated calculation using U_j^{E} and $\mathcal{I}_j^{\mathrm{E}}$ for J = 1. In this appendix, I continue the running example in this paper's main text and its Appendix A by working through J = 2 for the actual calculation. I focus on the calculation for assessing whether a specific covariate, j, violates the PH assumption. The end result will be the main text's Eq. (4).

Begin by recalling the expression for Therneau and Grambsch's PH test statistic (main text, Eq. (1); main text, fn. 6)

$$T_j^{act} = {U_j}^{\mathrm{E}} \mathcal{I}_j^{\mathrm{E}^{-1}} U_j^{\mathrm{E}'}$$

To calculate T_j^{act}, we will need two things: (1) the j-specific expanded score vector (${U_j}^{\mathrm{E}}$) and (2) the inverse of the j-specific expanded information matrix ($\mathcal{I}_j^{\mathrm{E}^{-1}}$).[32] I discuss both ${U_j}^{\mathrm{E}}$ and $\mathcal{I}_j^{\mathrm{E}}$ in the main text's Section II.

The j-specific expanded score matrix will be 1×3, equal to

$${U_j}^{\mathrm{E}} = \left[0 \quad 0 \sum\nolimits_k s_{j,k}\, g(t_k)\right]$$

where $s_{j,k}$ is the value of covariate j's unscaled Schoenfeld residual at time k. The j-specific expanded information matrix will be 3×3 and will equal (see Appendix A for further explanation)

$${\mathcal{I}_j}^{\mathrm{E}} = \begin{bmatrix} \sum_{k=1}^{K} \hat{V}(t_k, x_1) & \sum_{k=1}^{K} \widehat{\mathrm{Cov}}(t_k) & \sum_{k=1}^{K} \hat{V}(t_k, x_1) g(t_k) \\ \sum_{k=1}^{K} \widehat{\mathrm{Cov}}(t_k) & \sum_{k=1}^{K} \hat{V}(t_k, x_2) & \sum_{k=1}^{K} \widehat{\mathrm{Cov}}(t_k) g(t_k) \\ \sum_{k=1}^{K} \hat{V}(t_k, x_1) g(t_k) & \sum_{k=1}^{K} \widehat{\mathrm{Cov}}(t_k) g(t_k) & \sum_{k=1}^{K} \hat{V}(t_k, x_1) g^2(t_k) \end{bmatrix}$$

We need ${\mathcal{I}_j}^{\mathrm{E}}$'s inverse, $\mathcal{I}_j^{\mathrm{E}^{-1}}$, to compute T_j^{act}. The general formula for a matrix's inverse is

$$A^{-1} = \frac{1}{|A|} \mathrm{adj}(A)$$

[32] For succinctness in this appendix, I subsequently drop the "*act*" superscript when referring to these quantities.

where $|A|$ is A's determinant and $\mathrm{adj}(A)$ is A's adjoint matrix.

$\left|\mathcal{J}_j^{\mathrm{E}}\right|$ will end up being Eq. (4)'s denominator. This determinant will equal any row from $\mathcal{J}_j^{\mathrm{E}}$ multiplied by the same row, transposed, in $\mathcal{J}_j^{\mathrm{E}}$'s cofactor matrix. I mention this detail because, in the main text, I mentioned the complexity of the AC's denominator is one consequence of $\mathcal{J}_2 \neq 0$. Remembering the determinant formula makes clear one way in which $\mathcal{J}_2 \neq 0$ has this effect. To arbitrarily take $\mathcal{J}_j{}^{\mathrm{E}}$'s first row to illustrate, using placeholders for the actual cofactor matrix:

$$\left|\mathcal{J}_j^{\mathrm{E}}\right| = \begin{bmatrix} \sum_{k=1}^{K} \hat{V}(t_k, x_1) & \sum_{k=1}^{K} \widehat{\mathrm{Cov}}(t_k) & \sum_{k=1}^{K} \hat{V}(t_k, x_1) g(t_k) \end{bmatrix} \begin{bmatrix} \mathrm{cofactor}_{1,1} \\ \mathrm{cofactor}_{1,2} \\ \mathrm{cofactor}_{1,3} \end{bmatrix}$$

If $\mathcal{J}_2 = 0$, as it does for the approximation, the final element in $\mathcal{J}_j{}^{\mathrm{E}}$'s first row would be zero:

$$\begin{bmatrix} \sum_{k=1}^{K} \hat{V}(t_k, x_1) & \sum_{k=1}^{K} \widehat{\mathrm{Cov}}(t_k) & 0 \end{bmatrix} \begin{bmatrix} \mathrm{cofactor}_{1,1} \\ \mathrm{cofactor}_{1,2} \\ \mathrm{cofactor}_{1,3} \end{bmatrix}$$

Multiplying together these two vectors would make the third term drop, producing a determinant with two terms, at most.

I already mentioned $\left|\mathcal{J}_j^{\mathrm{E}}\right|$'s actual value in the main text (Eq. (5)). I will continue to refer to it as $\left|\mathcal{J}_j^{\mathrm{E}}\right|$, for brevity's sake.

$\mathcal{J}_j^{\mathrm{E}}$'s adjoint matrix is equal to the transpose of its cofactor matrix and its dimensions are identical to $\mathcal{J}_j^{\mathrm{E}}$. Similar to the determinant, $\mathrm{adj}\left(\mathcal{J}_j^{\mathrm{E}}\right)$ has many terms, to the point that its unsimplified form does not fit easily on a horizontally oriented letter-sized page. I have placed this matrix, in full, at the end of this appendix in a small font size. I will refer to its elements in text as $\mathrm{adj}\left(\mathcal{J}_j^{\mathrm{E}}\right)_{r,c}$ for the element in row *r*, column *c* and substitute in the expressions for any terms later.

We can now begin calculating T_j^{act}:

$$T_j^{act} = U_j{}^{\mathrm{E}} \mathcal{J}_j^{\mathrm{E}^{-1}} U_j^{\mathrm{E}\prime}$$

$$T_j^{act} = U_j{}^{\mathrm{E}} \left[\frac{1}{\left|\mathcal{J}_j^{\mathrm{E}}\right|} \mathrm{adj}\left(\mathcal{J}_j^{\mathrm{E}}\right)\right] U_j^{\mathrm{E}\prime}$$

$$T_j^{act} = \frac{1}{\left|\mathcal{J}_j^{\mathrm{E}}\right|} U_j{}^{\mathrm{E}} \mathrm{adj}\left(\mathcal{J}_j^{\mathrm{E}}\right) U_j^{\mathrm{E}'}$$

$1/\left|\mathcal{J}_j^{\mathrm{E}}\right|$ is a scalar, allowing it to be factored out from the to-be-multiplied matrices in such a way.

Multiplying the first two matrices together yields

$$U_j{}^{\mathrm{E}} \mathrm{adj}\left(\mathcal{J}_j^{\mathrm{E}}\right) = \left[0 \quad 0 \sum\nolimits_k s_{j,k}\, g(t_k)\right] \mathrm{adj}\left(\mathcal{J}_j^{\mathrm{E}}\right)$$

$$U_j{}^{\mathrm{E}} \mathrm{adj}\left(\mathcal{J}_j^{\mathrm{E}}\right) = \left[\mathrm{adj}\left(\mathcal{J}_j^{\mathrm{E}}\right)_{3,1} \sum\nolimits_k s_{j,k}\, g(t_k) \quad \mathrm{adj}\left(\mathcal{J}_j^{\mathrm{E}}\right)_{3,2} \sum\nolimits_k s_{j,k}\, g(t_k) \quad \mathrm{adj}\left(\mathcal{J}_j^{\mathrm{E}}\right)_{3,3} \sum\nolimits_k s_{j,k}\, g(t_k)\right]$$

Multiplying this product by $U_j^{\mathrm{E}'}$ produces

$$U_j{}^{\mathrm{E}} \mathrm{adj}\left(\mathcal{J}_j^{\mathrm{E}}\right) U_j^{\mathrm{E}'} = \left[\mathrm{adj}\left(\mathcal{J}_j^{\mathrm{E}}\right)_{3,1} \sum\nolimits_k s_{j,k}\, g(t_k) \quad \mathrm{adj}\left(\mathcal{J}_j^{\mathrm{E}}\right)_{3,2} \sum\nolimits_k s_{j,k}\, g(t_k) \quad \mathrm{adj}\left(\mathcal{J}_j^{\mathrm{E}}\right)_{3,3} \sum\nolimits_k s_{j,k}\, g(t_k)\right] \begin{bmatrix} 0 \\ 0 \\ \sum\nolimits_k s_{j,k}\, g(t_k) \end{bmatrix}$$

$$U_j{}^{\mathrm{E}} \mathrm{adj}\left(\mathcal{J}_j^{\mathrm{E}}\right) U_j^{\mathrm{E}'} = \mathrm{adj}\left(\mathcal{J}_j^{\mathrm{E}}\right)_{3,3} \left\{\sum\nolimits_k s_{j,k}\, g(t_k)\right\}^2$$

Finally, we divide by the determinant and insert the adjoint matrix's relevant element, producing

$$T_j^{act} = \frac{1}{\left|\mathcal{J}_j^{\mathrm{E}}\right|} U_j{}^{\mathrm{E}} \mathrm{adj}\left(\mathcal{J}_j^{\mathrm{E}}\right) U_j^{\mathrm{E}'} = \frac{\mathrm{adj}\left(\mathcal{J}_j^{\mathrm{E}}\right)_{3,3} \left\{\sum_k s_{j,k}\, g(t_k)\right\}^2}{\left|\mathcal{J}_j^{\mathrm{E}}\right|}$$

$$\boxed{T_j^{act} = \frac{\left\{\left[\sum_{k=1}^{K} \hat{V}(t_k, x_1) \sum_{k=1}^{K} \hat{V}(t_k, x_2)\right] - \left[\left(\sum_{k=1}^{K} \widehat{\mathrm{Cov}}(t_k)\right)^2\right]\right\} \left\{\sum_k s_{j,k}\, g(t_k)\right\}^2}{\left|\mathcal{J}_j^{\mathrm{E}}\right|}}$$

This expression matches the main text's Eq. (4), after we demean $g(t_k)$.

$$\mathrm{adj}\left(\mathcal{J}_j{}^{\mathrm{E}}\right)$$

$$= \begin{bmatrix} \left[\sum_{k=1}^{K}\hat{V}(t_k,x_2)\sum_{k=1}^{K}\hat{V}(t_k,x_1)g^2(t_k)\right]-\left[\left(\sum_{k=1}^{K}\widehat{\mathrm{Cov}}(t_k)g(t_k)\right)^2\right] & \left[\sum_{k=1}^{K}\hat{V}(t_k,x_1)g(t_k)\sum_{k=1}^{K}\widehat{\mathrm{Cov}}(t_k)g(t_k)\right]-\left[\sum_{k=1}^{K}\widehat{\mathrm{Cov}}(t_k)\sum_{k=1}^{K}\hat{V}(t_k,x_1)g^2(t_k)\right] & \left[\sum_{k=1}^{K}\widehat{\mathrm{Cov}}(t_k)\sum_{k=1}^{K}\widehat{\mathrm{Cov}}(t_k)g(t_k)\right]-\left[\sum_{k=1}^{K}\hat{V}(t_k,x_1)g(t_k)\sum_{k=1}^{K}\hat{V}(t_k,x_2)\right] \\ \left[\sum_{k=1}^{K}\widehat{\mathrm{Cov}}(t_k)g(t_k)\sum_{k=1}^{K}\hat{V}(t_k,x_1)g(t_k)\right]-\left[\sum_{k=1}^{K}\widehat{\mathrm{Cov}}(t_k)\sum_{k=1}^{K}\hat{V}(t_k,x_1)g^2(t_k)\right] & \left[\sum_{k=1}^{K}\hat{V}(t_k,x_1)\sum_{k=1}^{K}\hat{V}(t_k,x_1)g^2(t_k)\right]-\left[\left(\sum_{k=1}^{K}\hat{V}(t_k,x_1)g(t_k)\right)^2\right] & \left[\sum_{k=1}^{K}\hat{V}(t_k,x_1)g(t_k)\sum_{k=1}^{K}\widehat{\mathrm{Cov}}(t_k)\right]-\left[\sum_{k=1}^{K}\hat{V}(t_k,x_1)\sum_{k=1}^{K}\widehat{\mathrm{Cov}}(t_k)g(t_k)\right] \\ \left[\sum_{k=1}^{K}\widehat{\mathrm{Cov}}(t_k)\sum_{k=1}^{K}\widehat{\mathrm{Cov}}(t_k)g(t_k)\right]-\left[\sum_{k=1}^{K}\hat{V}(t_k,x_2)\sum_{k=1}^{K}\hat{V}(t_k,x_1)g(t_k)\right] & \left[\sum_{k=1}^{K}\widehat{\mathrm{Cov}}(t_k)\sum_{k=1}^{K}\hat{V}(t_k,x_1)g(t_k)\right]-\left[\sum_{k=1}^{K}\hat{V}(t_k,x_1)\sum_{k=1}^{K}\widehat{\mathrm{Cov}}(t_k)g(t_k)\right] & \left[\sum_{k=1}^{K}\hat{V}(t_k,x_1)\sum_{k=1}^{K}\hat{V}(t_k,x_2)\right]-\left[\left(\sum_{k=1}^{K}\widehat{\mathrm{Cov}}(t_k)\right)^2\right] \end{bmatrix}$$

Appendix C
Full List of Characteristics Varied in Simulations

- Three effect patterns for x_2 ("linear combination")

  In one pattern, x_2 has no main effect, to match Keele (2010) and Metzger (2023c), making the linear combination $XB = 0.001x_1 + 1x_2 \ln(t)$. In the two other patterns, x_2 has a main effect half the size of the time-varying effect (TVE). In one, the main effect and TVE are signed the same: $XB = 0.001x_1 + 0.5x_2 + 1x_2 \ln(t)$. In the other, they are signed oppositely: $XB = 0.001x_1 - 0.5x_2 + 1x_2 \ln(t)$.

- Three Weibull shape parameter (p) values

  {0.75, 1, 1.25}. See main text's Section III for discussion.

- Two sample sizes

  {100, 1000}. See main text's Section III for discussion.

- Two recorded duration types

  One of the types records the true continuous-time duration ("continuous"), and one coerces the continuous-time duration into an integer format by rounding up to the nearest integer ("coerced start-stop"), à la Metzger and Jones (2022). The second mimics a frequent way in which political science data records durations.

- Five levels of correlation between the two covariates

  {-0.65, -0.35, 0, 0.35, 0.65}. See main text's Section III for discussion.

- Two right-censoring (RC) patterns

  {random, top *rc*%}. See main text's Section III for discussion.

- Three right-censoring percentages (*rc*%)

  {0%, 25%, 50%}. See main text's Section III for discussion.

- Two means for x_1

  {0, 60}. Varying x_1's mean allows us to see whether the non-violator's magnitude impacts the calculations' performance.

- Two standard deviations (SDs) for x_1

  {1, 3}. Varying x_1's SD allows us to see whether the non-violator's dispersion impacts the calculations' performance, as we know a covariate's dispersion can affect its computed variance, in general.

Appendix D
Simulation Results: Other Characteristics

As a secondary concern, we can also evaluate how the calculations perform in response to the other factors I vary. In all the descriptions that follow, I hold all other characteristics constant and compare the characteristic of interest across comparable scenarios, the same as I did for the correlation scenarios (e.g., the 360 unique combinations formed by all the non-correlation-related characteristics). I use "performing better" as my statistic for the discussion, defined as the percentage of combinations in which a scenario associated with one value of the characteristic (e.g., Corr = 0) has either a lower false positive percentage (for x_1) or a higher true positive percentage (for x_2) compared to an otherwise identical scenario, except for that characteristic's value (e.g., Corr = 0.35). The percentages involve both x_1 and x_2, meaning every unique combination contributes two p-values to the calculations. If we were to redo Section IV's correlation calculations here, for instance, the percentage's denominator would be 720, from 360 unique combinations * 2 estimates' p-values.

TABLE 3. Calculation Performance: Other Characteristics

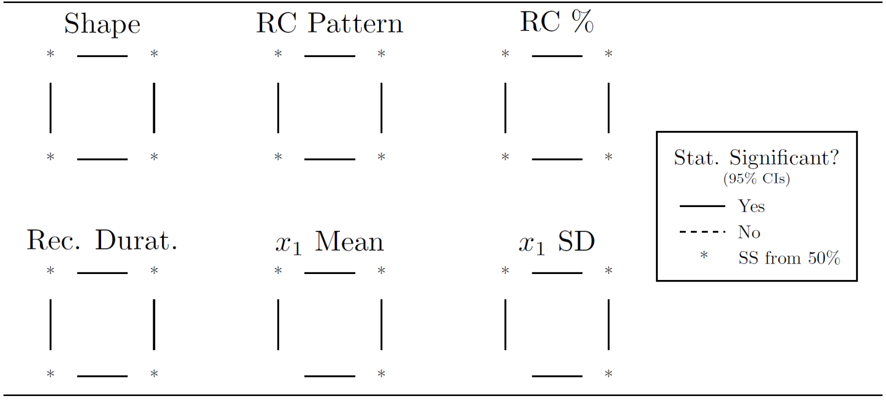
Shape: p = 0.75 Better?			RC Pattern: Random Better?			RC %: 0% Better?		
	n = 100	n = 1000		n = 100	n = 1000		n = 100	n = 1000
Approx.	68.25	27.25	Approx.	53.06	41.18	Approx.	69.58	28.75
	[65.53, 70.88]	[24.75, 29.86]		[50.44, 55.66]	[38.62, 43.77]		[67.13, 71.95]	[26.42, 31.16]
Actual	53.33	10.42	Actual	65.83	43.68	Actual	54.51	8.75
	[50.46, 56.19]	[8.75, 12.28]		[63.32, 68.28]	[41.10, 46.29]		[51.90, 57.11]	[7.34, 10.33]
Rec. Durat.: Continuous Better?			x_1 Mean: 60 Better?			x_1 SD: 3 Better?		
	n = 100	n = 1000		n = 100	n = 1000		n = 100	n = 1000
Approx.	54.28	19.89	Approx.	55.00	28.39	Approx.	47.28	27.06
	[51.94, 56.60]	[18.07, 21.81]		[52.67, 57.32]	[26.31, 30.53]		[44.95, 49.61]	[25.01, 29.17]
Actual	56.17	13.17	Actual	48.78	26.67	Actual	48.56	29.44
	[53.84, 58.48]	[11.64, 14.82]		[46.44, 51.12]	[24.64, 28.77]		[46.22, 50.89]	[27.35, 31.61]

In cells: % "yes, performed better," for comparable scenario–estimate combinations. Brackets: 95% analytic CIs. Global test excluded from all calculations.

For Table 3, I arbitrarily select one of the characteristic's values as the 'baseline' and tally the percentage of comparable scenario–estimate combinations where that baseline performs better. These numbers tell us about the calculations' performance in relative terms, for a given characteristic. If a characteristic does not impact performance, we would expect this percentage to be around 50%—between two comparable scenarios distinguished only by the characteristic's selected value, whether one scenario outperforms the other should come down to a coin flip.[33] This should rarely, if ever, be true, as I selected these characteristics precisely because they affect quantities appearing in the calculations' formulas.

For inference, I report each percentage's 95% analytic CIs in square brackets. Figure 4 provides a minimalist visual summary of Table 3—specifically, whether (a) each "perform better" percentage is statistically different from 50% (asterisks), (b) whether the calculations perform differently within a sample size (vertical lines), and (c) whether the same calculation performs differently across sample sizes (horizontal lines).[34]

FIGURE 4. Visual Summary: Statistically Significant Differences?



Based on Table 3.

Table 3 suggests all its constituent characteristics affect at least one of the calculations, evident in the way that 50% falls outside 22 of the 24 percentages' CIs, as Figure 4 also clearly shows. It also suggests all six characteristics affect the calculations differently in both sample sizes (Figure 4's solid vertical lines [fn. 34]). In all instances, a calculation's percentage for *n* = 100 is statistically distinguishable from its *n* = 1000 percentage (non-overlapping CIs within Table 3's panel rows/Figure 4's solid horizontal lines). Speaking in broad strokes:

[33] This value does not tell us about absolute performance—i.e., whether the calculations clear the relevant performance benchmark ($\hat{r}_p < 5\%$ (x_1) or $\geq 80\%$ (x_2)).

[34] Non-overlapping 95% CIs are sufficient to conclude two quantities are different, but overlapping CIs have two possible meanings (Austin and Hux 2002). For any overlapping CIs, I verify whether the quantities are different at the 95% level by computing the exact *p*-value from McNemar's test for paired binomial proportions (Rosner 2015, sec. 10.4); Figure 4 and the discussion are based on these checks.

- Weibull $h_0(t)$'s shape parameter (Table 3, top left): For both calculations, monotonically decreasing $h_0(t)$s perform better for n = 100 than flat or increasing $h_0(t)$s (percentages > 50%), but not for n = 1000 (< 50%).
- RC pattern (Table 3, top middle): The AC performs better for both sample sizes when the censoring pattern is random (highest within-column percentage). The calculations also perform differently for both sample sizes.
- RC percentage (Table 3, top right): Similar to the RC pattern, the RC percentage matters more in smaller sample sizes, where RC % = 0 does better than the comparable RC % ≠ 0 scenario (percentage > 50%). The approximation is more sensitive to right censoring in both sample sizes (highest within-column percentage).
- Recorded duration type (Table 3, bottom left): Which type performs better depends on sample size. Additionally, the two calculations' percentages behave differently for n = 100 and n = 1000.
- x_1's mean (Table 3, bottom middle): x_1's mean impacts the approximation when sample sizes are small, but not the AC. When n = 1000, x_1's mean matters more for both calculations, with the *smaller* mean (= 0) outperforming the larger mean. In addition, the two calculations perform differently for both sample sizes.
- x_1's dispersion (Table 3, bottom right): x_1's dispersion has similar patterns as x_1's mean in all respects.

Appendix E
Effect of Including Irrelevant Time Interactions: Cox Model

From a statistical perspective, including an interactive effect when none exists has the same implications for the Cox model as it does for any regression model: the model will be inefficient, relative to a correctly specified model. More specifically, the model will be underpowered, because interaction terms effectively divide the sample into smaller subsamples, and less data means less precise estimates, all else equal (Gelman, Hill, and Vehtari 2020, 301). Substantively, then, we become more prone to making Type II errors, all else equal.

To verify this behavior, I reran the set of scenarios from Figure 1 in which the Weibull's baseline hazard is flat (p = 1, center column). I add to my current simulations by also estimating the correct model, per the PH test's output, and record the estimated coefficients from that model. Each simulation draw returns two sets of "correct" model results: the correct model specification according to the approximation, and the correct model specification according to the AC, which I refer to as "___________"'s final specification in my discussion.

I classified each draw as falling into one of four mutually exclusive and exhaustive categories:

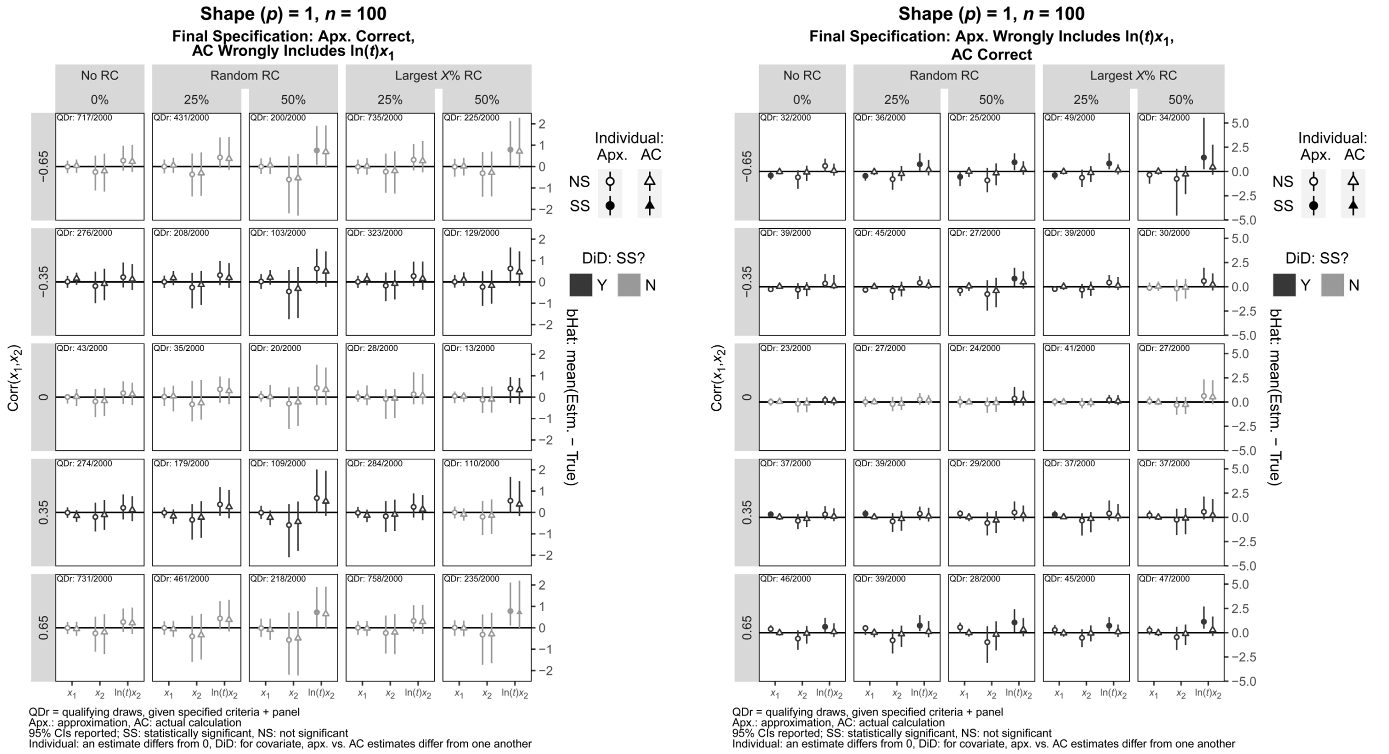
1. The final specifications for both calculations match the true DGP, meaning both calculations correctly classify x_1 as a non-PH violator and x_2 as a PH violator.
2. The approximation's final specification matches the true DGP, but the AC's does not.
3. The approximation's final specification does not match the true DGP, but the AC's does.
4. Neither calculation's final specification matches the true DGP.

I further subdivide the middle two categories, because there are multiple ways in which a final specification could be incorrect. It could (a) improperly classify x_1 as a PH violator, (b) improperly classify x_2 as a non-PH violator, or (c) both (a) and (b). Situation (a) means we would have an unnecessary interaction term between $g(t)$ and x_1, but a necessary interaction term between $g(t)$ and x_2. This situation has the highest number of estimated parameters, across the three ways in which a final specification could be incorrect. It is also the situation of most interest, because we can juxtapose the performance of the incorrectly specified final model from one calculations with the correctly specified model's performance for the other calculation, for the same subset of simulation draws.

I report information about estimate bias (Figure 5) and precision (Figure 6) for Categories 2a (left panel) and 3a (right panel). At each subpanel's upper-left corner, I provide information about how many of the scenario's total draws fall into this particular category. For instance: for the scenario in which no RC exists and the covariates are correlated at -0.65 (first row, first column), the approximation's specification is entirely correct and the AC's is incorrect by including ln(t)x_1 in 717 of the scenario's 2000 draws (left panel, either figure). The reverse is true less frequently. The approximation's specification incorrectly includes ln(t)x_1 and the AC's is correct in 32 of the scenario's 2000 draws (right panel, either figure).

A. Unbiasedness

FIGURE 5. PH-Corrected Simulation Results: Estimates' Unbiasedness



DGP = Figure 1 and 2's DGP, center column.

We should see unbiased coefficients if the Cox model behaves similarly to other regression models with respect to the effect of including unnecessary interactions. I use hollow symbols in Figure 5 to denote unbiased estimates, defined as a situation in which the mean difference between the parameter's true value and its estimated value is not statistically distinguishable from zero based on that estimated difference's 95% percentile-based CI, and I use shaded symbols to denote biased estimates. In line with those expectations, nearly all the estimates are unbiased in both of Figure 5's panels.

For Figure 5's left panel, all but five of the 150 estimates are unbiased. All five occur for ln(t)x_2, |Corr| = 0.65, and *rc*% = 50%. The approximation's estimates (circles) are always biased for this combination of characteristics, regardless of RC pattern, whereas the AC's estimates (triangles) are only biased when the RC pattern is top *X*%. On the face of it, this could suggest including an unnecessary interaction term has a low cost: the AC's incorrect specification is unbiased for 74 of 75 estimates in Figure 5's left panel, while the approximation's correct specification produces biased estimates for x_2's time-varying effect when the data are heavily censored and the correlation between violator and non-violator is high.

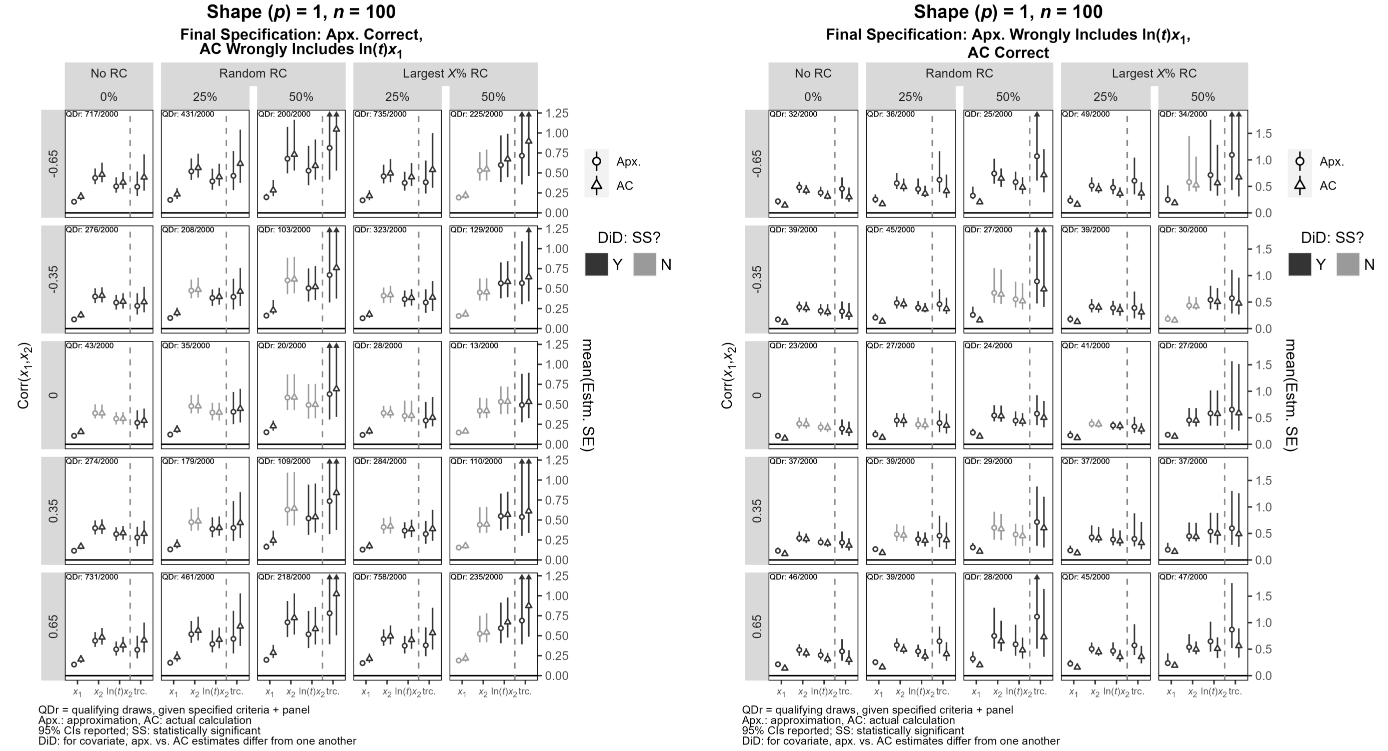
However, Figure 5's right panel casts some doubt on this story. In this panel, the AC's final specification is correct, but the approximation's final specification incorrectly includes the ln(t)x_1 interaction.[35] Seventeen of the 150 estimates are biased in this panel. If including the interaction term was always a costless hedge, we would expect to see some of the AC's final, fully correct specification estimates be biased in this panel, similar to how some of the approximation's estimates were biased in Figure 5's left panel when its final specification was fully correct. This does not occur. All 17 biased estimates from the right panel are exclusively from the approximation's (incorrect) final specification. In addition, the approximation's four biased estimates from the left panel are present in the set of 17 biased estimate from the right panel, suggesting that something about how the approximation performs in those scenarios brings about biased estimates of x_2's time-varying effect, regardless of whether the model's final specification is correct.

[35] Few draws fall into this category: there are approximately 20–50 draws, constituting ~1%– 2.5% of each scenario's 2000 draws.

On the whole, Figure 5's evidence is broadly in line with our expectations: including an unnecessary interaction term does not induce biased estimates in absolute terms, generally speaking.

B. Relative Efficiency

FIGURE 6. PH-Corrected Simulation Results: Estimates' Relative Efficiency



Shape (p) = 1, n = 100
Final Specification: Apx. Correct, AC Wrongly Includes $\ln(t)x_1$

Shape (p) = 1, n = 100
Final Specification: Apx. Wrongly Includes $\ln(t)x_1$, AC Correct

QDr = qualifying draws, given specified criteria + panel
Apx.: approximation, AC: actual calculation
95% CIs reported; SS: statistically significant
DiD: for covariate, apx. vs. AC estimates differ from one another

QDr = qualifying draws, given specified criteria + panel
Apx.: approximation, AC: actual calculation
95% CIs reported; SS: statistically significant
DiD: for covariate, apx. vs. AC estimates differ from one another

DGP = Figure 1 and 2's DGP, center column.

We should also see that the model specification with an unnecessary interaction term is less efficient, in relative terms, than the correct specification. To formally assess relative efficiency, I take the variance-covariance matrix's trace in every draw and compute the average across all the draws (labeled "trc." in Figure 6). A multivariate estimator X is more efficient than an alternative Y if the trace of X's variance-covariance matrix is smaller than the trace of Y's variance-covariance matrix (Kennedy 2003, 38–39).

I report the trace for each scenario in Figure 6, at the far right of each facet. Dark-gray shading indicates a statistically significant difference between the approximation's and AC's trace estimate for that scenario. As expected, the correct specification does indeed have a smaller estimated trace compared to the incorrect specification, in all 50 of Figure 6's facets. More specifically, the correctly specified approximation’s estimated trace is smaller in Figure 6’s left panel than that of the incorrectly specified AC, and in Figure 6’s right panel, the correctly specified AC’s estimated trace is smaller than the incorrectly specified approximation’s estimated trace. This difference is also statistically distinguishable from zero in all 50 facets (all trace estimates are dark gray). In sum, Figure 6 is also in line with our general expectations: specifications with the unnecessary interaction are less efficient than those without it.

Appendix F
Agerberg and Kreft (2020): Additional Details

A. Research Design Details

Agerberg and Kreft (2020, hereafter "A&K") use a Cox model with standard errors clustered on country assess their two hypotheses, with their estimation sample comprising 136 countries from 1990 to 2013.[36] A&K declare all countries at risk of adopting a national legislative gender quota in 1990, "the start of the global wave of gender quota adoption" (A&K 2020, 301). However, 10 countries in their sample do not exist in 1990. I adjust their coding and consider these countries at risk only once they become independent. All countries exit the sample once they have adopted a gender quota.

A&K's covariate of interest is based on any civil conflicts occurring within a country between 1990 and 2013 (A&K 2020, 301–2). Their main independent variable coded 1 for all of country X's conflict and post-conflict years if there are reports of widespread or systematic occurrences of SV during X's civil conflict ("high SV conflicts" → "HSVC") and 0 otherwise. A&K also include a variable coded 1 for all of X's conflict and post-conflict years if there were no or isolated reports of SV during X's civil conflict ("low SV conflicts" → "LSVC"), making the omitted category countries that never experienced any civil conflicts between 1990 and 2013. In addition to these two variables, A&K's main specification includes eleven additional control variables (their Table 1, Model 2).

B. Model Results

See next page.

[36] Their sample "exclude[s] long-term consolidated and developed democracies" (A&K 2020, 301).

TABLE 4. Agerberg and Kreft (2020): Cox Models with PH Violation Corrections

	A&K Original*	Viols. Diagnosed Using:	
		Appx.	AC
LSVC	-0.854	-0.961	-0.904
	(0.834)	(1.002)	(0.840)
HSVC	1.224	1.410	-0.144
	(0.742)	(0.930)	(2.139)
GDPPC (ln)	0.009	0.012	0.025
	(0.143)	(0.142)	(0.147)
Polity	0.024	0.003	0.033
	(0.068)	(0.070)	(0.071)
Conflict intensity: Low	-0.399	-4.062	-0.293
	(0.762)	(1.677)	(3.570)
Conflict intensity: High	-0.798	-3.747	-0.701
	(0.771)	(1.397)	(5.287)
Peacekeeping operation	0.854	1.155	0.897
	(0.520)	(0.541)	(0.530)
Foreign aid (ln)	0.184	0.135	0.174
	(0.092)	(0.093)	(0.095)
Regional quota diffusion	0.017	0.017	0.019
	(0.011)	(0.012)	(0.011)
Islamic heritage	0.170	0.196	0.636
	(0.334)	(0.336)	(1.945)
Women's civil liberties (1990)	-0.299	-0.472	-0.447
	(0.761)	(0.804)	(0.775)
Electoral system: PR	1.222	1.378	1.309
	(0.365)	(0.381)	(0.374)
Electoral system: Mixed	0.847	0.898	0.890
	(0.395)	(0.398)	(0.400)
t * (Conflict intensity: Low)		0.233	-0.003
		(0.084)	(0.278)
t * (Conflict intensity: High)		0.177	-0.018
		(0.061)	(0.856)
t * HSVC			0.173
			(0.247)
t * Polity			-0.020
			(0.023)
t * (Islamic heritage)			-0.082
			(0.360)
t * (Women's civil liberties (1990))			0.009
			(0.005)
n_{Fail}	57	57	57
n	2,642	2,642	2,642

* With corrected duration coding for 10 countries

Hazard coefficients reported. Standard errors clustered by country in parenthesis. Models estimated using Efron tie correction.

Appendix G
Meta-Analysis Details[37]

A. Search Procedure

To identify the situations practitioners generally encounter when using Cox models, I examined all substantive articles published in eight journals from August 2019 to the present, including any FirstView/forthcoming articles available online as of January 15, 2023:

- Generalist journals
    - *American Journal of Political Science*
    - *American Political Science Review*
    - *Journal of Politics*
- International relations (IR) journals
    - General
        - *International Organization*
        - *International Studies Quarterly*
    - Security
        - *Conflict Management and Peace Science*
        - *Journal of Conflict Resolution*
        - *Journal of Peace Research*

I selected several IR journals because of the frequency with which international security scholars in particular employ Cox models. I started the search in August 2019 because `survival` 3.0-10, the first release in which the PH test's default is the AC, was released via GitHub in this month.

I used both JSTOR and individual publisher sites to search for "Cox proportional hazard model," "hazard model," and "Cox semi-parametric model" and looked through the search results to confirm whether an article used a Cox model. For the articles that did, I then looked for the article's replication

[37] I thank Nicholas Bush for his research assistance.

materials in the journal's centralized data repository and, if not present there, authors' individual webpages.

The search yielded 55 articles in total. Of these, 41 reported a Cox model in the main text, which I defined as the main text having either (a) a table with Cox model estimates or (b) a figure displaying predicted quantities from a Cox model. Twenty-eight of these 41 articles have publicly available replication data.

B. Covariate Correlations

Determining how many political science applications have correlated PH violators and non-violators helps speak to how frequently we may encounter correlation-induced false positives with the AC in practice. I could include 27 of the 28 articles I mentioned above for this portion of the analysis.[38] I selected the first Cox model reported with control variables from each of these 27 papers and checked for PH violations using $g(t) = \ln(t)$.[39,40] There is no clear-cut choice as to which calculation to use for the check because we know both calculations have problematic behavior in certain circumstances. I opted for the approximated PH test, as we know it is (1) insensitive to correlations between violators and non-violators, relative to the AC, and (2) generally (though not always) less prone to false positives. I used a conservative PH test *p*-value threshold and classified any covariate with $p < 0.1$ as a violator.[41] Twenty-four of the 27 articles had at least one PH violator, according to these $g(t)$ and *p*-value criteria.

[38] The excluded paper has a conditional frailty model (Blair, Grossman, and Weinstein 2022), from which Stata is incapable of estimating the PH test. The model also would not estimate in R using `survival` 2.44.1, the last release with the approximated test, and `survival` 2.44.1 does not permit Cox model results from a later `survival` release to be used with 2.44.1's PH test routine.

[39] Choosing $g(t)$ should be motivated by specific features of the application's dataset, in a real analysis (Park and Hendry 2015). Wielding such specialized knowledge is beyond the scope of my analysis here. I selected $g(t) = \ln(t)$, as political science applications frequently use this transformation, for better or worse. I also checked $g(t) =$ rank (Table 6, end of this appendix) and $g(t) = t$ (Table 7, end of this appendix), and the patterns for both were broadly similar to $\ln(t)$.

[40] I omitted any interactions correcting for PH violations from the specification in my reanalysis, if any such interactions were present.

[41] For the three articles with stratified hazards, I appropriately adjust the PH test to account for the strata (Metzger and Jones 2021). Additionally, two of the three with-strata articles estimate collapsed covariate effects. I classify a covariate as a PH violator if it meets the *p*-value criterion in any stratum for those two articles.

Next, I examined the pairwise correlations between every PH violator and non-violator in each article's estimation sample. Every covariate was classified as a violator in one article, leaving 23 articles with at least one violator and at least one non-violator. I take the correlations' absolute value for simplicity in reporting because the main simulations show only the correlation's magnitude affects the AC's false positive rate, not the correlation's sign.

The meta-analysis shows political scientist applications frequently have PH violators correlated with non-violators (Table 5). Twenty of the 23 articles (86.96%) have at least one PH violator moderately correlated with at least one non-PH violator. Furthermore, these 20 articles do not simply have one pairing satisfying this criterion, but several: the median number of pairings (3.5) implies ten of the 20 articles have four or more violator–non-violator pairings with |Corr| ∈ (0.25, 0.5] (mean: 5.15 pairings). Fewer articles (9 of 23) have highly correlated (PH violator)–(non-PH violator) pairings.[42]

TABLE 5. Meta-Analysis: Number of Articles Meeting Correlation Criteria ($g(t) = \ln(t)$, p-value ≤ 0.1)

|Corr| Range	*Frequency*	*Central Tendency (Mn./Med.)*	
		ALL ARTICLES	WITHIN CATEGORY
(0, 0.25]	23	21.913	21.913
	(100.0%)	19	19
(0.25, 0.5]	20	4.478	5.150
	(86.96%)	3	3.5
(0.5, 0.75]	9	1	2.556
	(39.13%)	0	3
(0.75, 1]	3	0.130	1
	(13.04%)	0	1

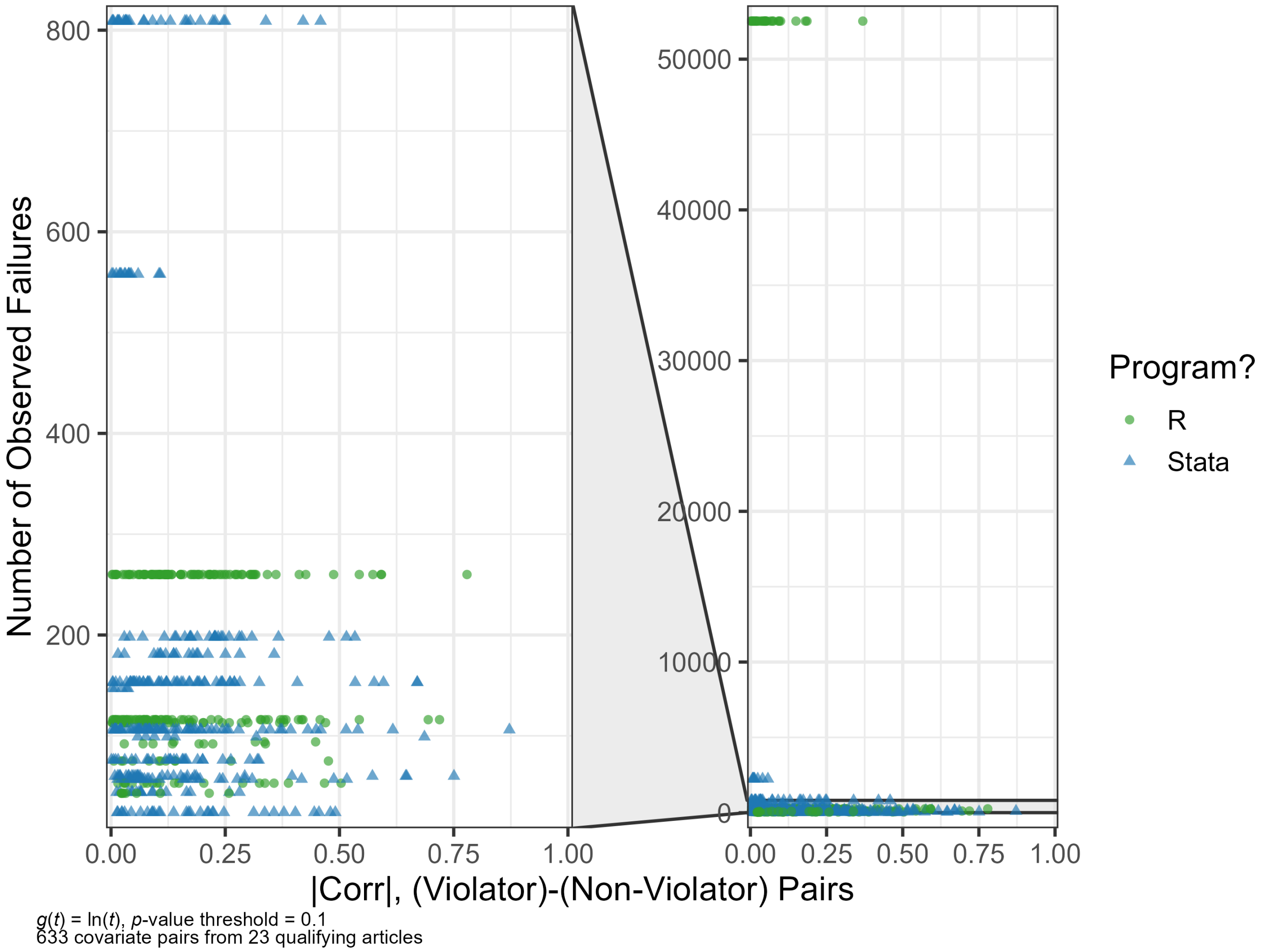
# qualifying articles = 23 with $g(t) = \ln(t)$ and PH test p-value threshold = 0.1.
Frequency: Articles with at least one (violator)–(non-violator) pairing whose correlation's absolute value falls into the specified range.
Central Tendency: The average/median number of (violator)–(non-violator) pairings that fall into the specified range among qualifying articles ("qualifying" criterion = *subcolumn headers*). First cell in vertical pair: mean; second cell: median.

Figure 7 provides a visual representation of the correlations for all (PH violator)–(non-PH violator) pairings, plotted against the number of observed events in the model where the covariate pair appears. I

[42] The (0.75,1] set's three articles are also present in the (0.5,0.75] set, yielding 9 unique articles where the correlation for at least one PH violator–non-violator pairing falls in the range (0.5,1]. The nine unique articles from (0.5,1] also all appear in the (0.25, 0.5] set, meaning there are 20 unique articles that have at least one pairing with |Corr| > 0.25.

use different shapes and colors to denote articles using R (green circles) vs. Stata (blue triangles) to estimate its Cox model.[43]

FIGURE 7. Meta-Analysis: Number of Failures vs. Covariate Correlations



C. AC Use Theoretically Warranted

In addition, I examine the number of articles where using the AC might be warranted, in theory—situations in which the approximation's simplifying assumption may not hold. The approximation assumes $\hat{V}(t_k)$ is constant across all t_k (main text, Sect. II.B), which we can synonymously state as $\hat{V}(t_k)$

[43] The article with over 50000 failure events examines Ecuadorian legislators' speech length in 11 legislative terms, covering 30 calendar years (Vallejo Vera and Gómez Vidal 2022). The article with the fewest number of events, 14, examines creating parliamentary organs within regional international organizations (Lenz, Burilkov, and Viola 2019).

being constant across all risksets. Any situation in which multiple risksets exist has the potential to violate this assumption (Therneau and Grambsch 2000, 141). At minimum, all estimated Cox models have as many risksets as there are unique failure times in the overall estimation sample. However, Cox models with strata can have more risksets than just this minimum amount, and the reasons motivating researchers to include strata could also imply $\hat{V}(t_k)$'s value differs across risksets (Therneau and Grambsch 2000, 141–42). For this reason (among others), stratified Cox models have attracted the most concern about violating the approximation's simplifying assumption, traditionally (Therneau and Grambsch 2000, 141; main text, fns. 1 and 5).

I again examine the first Cox model with controls reported in each article. I count how many of these models have strata, among the original 28 articles employing a Cox model in the main text for which I have replication data. Only four do (14.3%).[44] When paired with this number, the previous subsection's ~87% is even more striking. It suggests only a small portion of political science applications are potentially most affected by the issues motivating the switch to the AC, but a far larger proportion of applications are characterized by situations associated with the AC's high false positive rate.

[44] This number is likely a slight-to-modest underestimate due to the meta-analysis' search criteria, which would exclude any article describing its estimated models *exclusively* in terms of "competing risks" (and then estimating a cause-specific hazard via the Cox, not a subdistribution hazard [Fine–Gray's competing-risks model]) or "repeating events", without "Cox" being mentioned once anywhere. Strata are a signature feature of competing risks models and also appear in some repeated events models.

D. Supplemental Tables: Alternative g(t) *Choices*

TABLE 6. Meta-Analysis: Number of Articles Meeting Correlation Criteria (*g*(*t*) = rank(*t*), *p*-value ≤ 0.1)

|Corr| Range	*Frequency*	*Central Tendency (Mn./Med.)*	
		ALL ARTICLES	WITHIN CATEGORY
(0, 0.25]	24 (100.0%)	22.542 18	22.542 18
(0.25, 0.5]	20 (83.33%)	4.417 3	5.300 3.5
(0.5, 0.75]	9 (37.50%)	1 0	2.667 2
(0.75, 1]	2 (8.33%)	0.125 0	1.5 1.5

# qualifying articles = 24 with *g*(*t*) = rank and PH test *p*-value threshold = 0.1.

Frequency: Articles with at least one (violator)–(non-violator) pairing whose correlation's absolute value falls into the specified range.

Central Tendency: The average/median number of (violator)–(non-violator) pairings that fall into the specified range among qualifying articles ("qualifying" criterion = *subcolumn headers*). First cell in vertical pair: mean; second cell: median.

TABLE 7. Meta-Analysis: Number of Articles Meeting Correlation Criteria (*g*(*t*) = *t*, *p*-value ≤ 0.1)

|Corr| Range	*Frequency*	*Central Tendency (Mn./Med.)*	
		ALL ARTICLES	WITHIN CATEGORY
(0, 0.25]	22 (100.0%)	23.636 16	23.636 16
(0.25, 0.5]	19 (86.36%)	4.409 3	5.105 3
(0.5, 0.75]	9 (40.91%)	0.864 0	2.111 2
(0.75, 1]	1 (4.55%)	0.091 0	2 2

# qualifying articles = 22 with *g*(*t*) = *t* and PH test *p*-value threshold = 0.1.

Frequency: Articles with at least one (violator)–(non-violator) pairing whose correlation's absolute value falls into the specified range.

Central Tendency: The average/median number of (violator)–(non-violator) pairings that fall into the specified range among qualifying articles ("qualifying" criterion = *subcolumn headers*). First cell in vertical pair: mean; second cell: median.

Appendices Works Cited

Agerberg, M., and A.-K. Kreft. 2020. "Gendered Conflict, Gendered Outcomes: The Politicization of Sexual Violence and Quota Adoption." *Journal of Conflict Resolution* 64 (2–3): 290–317.

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

Blair, C. W., G. Grossman, and J. M. Weinstein. 2022. "Forced Displacement and Asylum Policy in the Developing World." *International Organization* 76 (2): 337–78.

Gelman, A., J. Hill, and A. Vehtari. 2020. *Regression and Other Stories*. Cambridge: Cambridge University Press.

Keele, L. 2010. "Proportionally Difficult: Testing for Nonproportional Hazards in Cox Models." *Political Analysis* 18 (2): 189–205.

Kennedy, P. 2003. *A Guide to Econometrics*. 5th ed. Cambridge, MA: MIT Press.

Lenz, Tobias, Alexandr Burilkov, and Lora Anne Viola. 2019. "Legitimacy and the Cognitive Sources of International Institutional Change: The Case of Regional Parliamentarization." *International Studies Quarterly* 63 (4): 1094–1107.

Metzger, S. K. 2023. "Proportionally Less Difficult?: Reevaluating Keele's 'Proportionally Difficult.'" *Political Analysis* 31 (1): 156–63.

Metzger, S. K., and B. T. Jones. 2021. "Properly Calculating `estat phtest` in the Presence of Stratified Hazards." *Stata Journal* 21 (4): 1028–33.

------. 2022. "Getting Time Right: Using Cox Models and Probabilities to Interpret Binary Panel Data." *Political Analysis* 30 (2): 151–66.

Park, S., and D. 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.

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

Therneau, T. M. 2021. "`cox.zph`: `zph.Rnw` Documentation." https://github.com/therneau/survival/blob/f2567b77252ac7935eba0ead364665c654ef28d3/noweb/zph.Rnw.

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

Vallejo Vera, S., and A. Gómez Vidal. 2022. "The Politics of Interruptions: Gendered Disruptions of Legislative Speeches." *Journal of Politics* 84 (3): 1384–1402.