SUPPLEMENTARY MATERIAL

How to analyze seed germination data using statistical time-to-event analysis: nonparametric and semiparametric methods

James N. McNair[1], Anusha Sunkara[2], and Daniel Frobish[2]

[1]Annis Water Resources Institute, Grand Valley State University, [2]Department of Statistics, Grand Valley State University

Here we provide the following supplementary material: (1) a derivation of the life-table estimator of the survivor function, (2) an overview of the Mantel-Haenzel test for comparing groups based on interval data, (3) some additional statistical details about exact-data methods for comparing groups, (4) an assessment of the effect of random seed loss, (5) a derivation of the recommended graphical procedure for assessing the proportional hazards (PH) assumption on which the Cox model is based, (6) a detailed exposition of the steps involved in applying the Cox model to the Japanese knotweed test data, and (7) sample R and SAS code for implementing the various methods of time-to-event analysis discussed in the text.

The life-table estimator of the survivor function

The basic life-table estimator given by equation (5) of the text can be derived as follows. Let the initial number of seeds be N. Given observation times $0 = a_0 < a_1 < a_2 < \cdots < a_m < \infty$ (chosen in advance), let the intervals between observations be $I_j = (a_{j-1}, a_j]$ for $j = 1, 2, 3, \ldots, m$. We assume the I_j are long enough relative to the rate at which germination events occur so that multiple events commonly occur within an interval. Let D_j be the number of germination events occurring within interval I_j. We assume the D_j are known but the exact event times are unknown. The number D_j of events occurring in interval $I_j = (a_{j-1}, a_j]$ is not known until observation time a_j. Seed losses also may occur, due, for example, to accidents while handling seeds on observation days. We assume the cumulative number W_j of losses within each interval I_j is known but the exact times at which they occur are unknown.

Let q_j denote the probability that a particular seed germinates during interval I_j, given that it did not germinate previously. Then $p_j = 1 - q_j$ is the probability that the seed fails to germinate (i.e., survives the germination process) during interval I_j, given that it did not germinate previously. In other words, p_j is the probability that the germination time is greater than a_j, given that it is greater than a_{j-1}. Recalling the definition of the survivor function, we must have $S(a_0)=S(0)=1,$ while for larger a_j the elementary rules of conditional probability tell us that

$$S(a_j) = S(a_{j-1})\, p_j, \quad j = 1, 2, 3, \ldots, m. \tag{S.1}$$

Applying this relationship iteratively to times a_{j-1}, a_{j-2}, and so on, it follows that

$$S(a_j) = S(a_{j-1})\, p_j = S(a_{j-2})\, p_{j-1}\, p_j = \cdots = \prod_{i=1}^{j} p_i = \prod_{i=1}^{j} (1-q_i). \tag{S.2}$$

If no seeds are lost before the end of an experiment, it is reasonable to estimate q_j by $\hat{q}_i = D_i / N_i$, where N_i is the number of seeds at risk of germination at the beginning of interval I_i and $N_1 = N$. The resulting estimator $\hat{S}(a_j)$ for $S(a_j)$ is then

$$\hat{S}(a_j) = \prod_{i=1}^{j} (1 - D_i / N_i), \tag{S.3}$$

which is equation (5) of the text. This is the standard life-table estimator of the survivor function at observation times a_i when no seed loss occurs. The usual modification to account for seed loss is outlined in the text.

Equation (S.3), and equations (5), (6), and (7) of the text, tell us how to estimate the survivor function at the observation times, but how do we estimate its value at intermediate times? Recalling that the observation times are intentionally chosen to ensure that multiple events occur in many of the intervals I_i, we do not want to assume that the germination process stops between successive observation times. But this would be the implication of assuming the survivor function is a step function that changes only at the observation times. (Note: the Kaplan-Meier estimator yields a step function, but it assumes the data are exact, implying that no germination events occur between recorded germination times.) It is more reasonable to assume events occur more or less uniformly during each interval I_i, and therefore to estimate $S(t)$ for $a_i < t < a_{i+1}$ simply by linearly interpolating between $\hat{S}(a_i)$ and $\hat{S}(a_{i+1})$. That is,

$$\hat{S}(t) = \varphi(t)\, \hat{S}(a_{i+1}) + [1 - \varphi(t)]\, \hat{S}(a_i) = [1 - \varphi(t) \hat{q}_{i+1}]\, \hat{S}(a_i), \quad a_i < t < a_{i+1}, \tag{S.4}$$

where

$$\varphi(t) = (t - a_i) / (a_{i+1} - a_i) \tag{S.5}$$

and we used the fact that $\hat{S}(a_{i+1}) = \hat{S}(a_i)(1 - \hat{q}_{i+1})$. These equations can be used to calculate the median and other quantiles of the germination time distribution. For example, to find the median, we find the pair of survivor-function values $\hat{S}(a_i)$ and $\hat{S}(a_{i+1})$ that bracket 0.5. The corresponding observation times a_i and a_{i+1} then bracket the median. Substituting these values in equations (S.4) and (S.5), setting $\hat{S}(t) = 0.5$, and solving for t yields an estimate of the median.

Comparing groups using interval data

A useful discussion of methods for comparing groups based on interval data is provided by Elandt-Johnson and Johnson (1980). Though now somewhat dated, this is the only text we know that presents group-comparison methods for interval data; all other texts restrict attention to methods for exact data. As in the case of estimators for the survivor function, however, the two types of methods can be expected to yield results that very similar if little or no seed loss occurs (see below).

Suppose there are $K \geq 2$ groups, and let us initially assume there are no seed losses. Let $0 = a_0 < a_1 < a_2 < \cdots < a_m < \infty$ be the observation times, and let $I_i = (a_{i-1}, a_i]$ for $i = 1, 2, 3, \ldots, m$ be the intervals between observations. Let D_{ij} be the number of germination events in interval I_i in group j, let N_{ij} be the number of seeds at risk in group j at the beginning of interval I_i, and let $D_i = \sum_j D_{ij}$ and $N_i = \sum_j N_{ij}$. Mantel and Haenszel (1959) noted that if the total number D_i of germination events in all groups and the number N_{ij} at risk in each group j are regarded as fixed

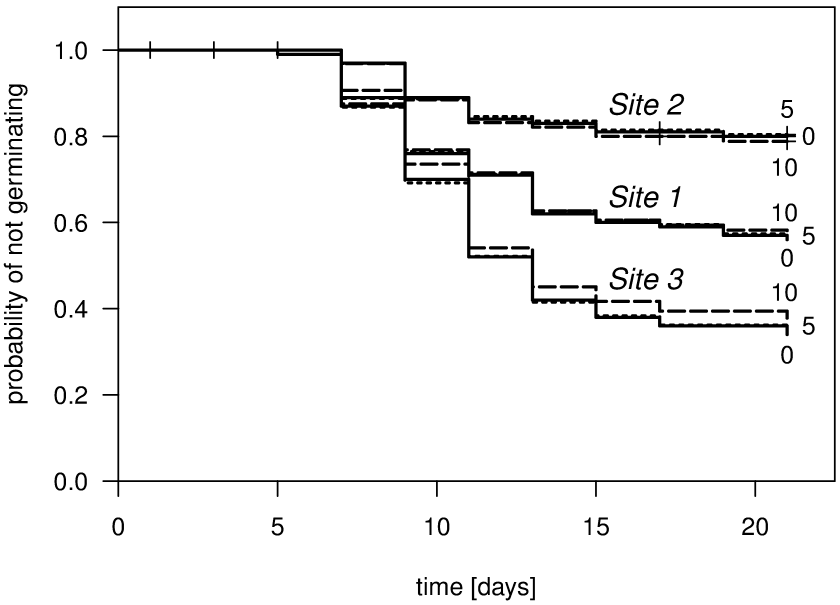

Figure S.1. The effect of small amounts of random seed loss on Kaplan-Meier estimates of the survivor function. The original data are for Japanese knotweed seeds collected from all three study sites on date 3 and are the same as in panels A and C of Figure 5 of the text. Three Kaplan-Meier survivor functions (labeled 0, 5, and 10) are shown for each study site: one for 0 lost seeds (original data), and one each for 5 and 10 artificially created seed losses (see text for details).

for each interval I_i, then we have a sequence of m contingency tables, each of size $K \times 2$ (groups $\times$ germination status) with fixed row and column totals. Under the null hypothesis that the number of germination events in each time interval is independent of group, the expected number of germination events during interval I_i in group j is $N_{ij}(D_i/N_i)$. Therefore, the sum Z_j of the differences between the observed (D_{ij}) and expected number of events over all time intervals in group j is given by

$$Z_j = \sum_{i=1}^{m} \left[D_{ij} - N_{ij}\left(D_i/N_i\right) \right]. \tag{S.6}$$

Under the null hypothesis, the number of germination events in any $K - 1$ groups will follow a multivariate (or univariate, if $K = 2$) hypergeometric distribution, from which the variances and covariances of the Z_j can be readily determined (e.g., Johnson et al., 1997). Let $\mathbf{z}$ be a vector of any $K - 1$ of the Z_j, and let $\boldsymbol{\Sigma}^{-1}$ be the inverse of the variance-covariance matrix. (Note: the Z_j sum to 1, so any $K - 1$ values completely determine the K-th.) Then for sufficiently large samples, $\mathbf{z}$ will have an approximately multivariate normal distribution, and the quadratic form Q given by

$$Q = \mathbf{z}^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \mathbf{z} \tag{S.7}$$

therefore will be distributed approximately as chi squared with $K - 1$ degrees of freedom. This statistic is the basis for the Mantel-Haenszel test for homogeneity ($K > 2$) or pair-wise group differences ($K = 2$). It is best viewed as a test for differences in hazard rate, because the number of seeds at risk in each group and time interval is regarded as fixed.

If seed losses occur during the experiment, they can be handled in the same way as in the life-table estimator for the survivor function (Elandt-Johnson and Johnson, 1980). Thus, we replace N_{ij} with the adjusted or effective number N_{ij}' of seeds at risk. The usual choice of N_{ij}' is $N_{ij} - 0.5W_{ij}$, where W_{ij} is the number of seeds lost during interval I_i in group j.

As noted in the text, neither R nor SAS currently provides interval-data methods for comparing life-table survivor functions as part of their built-in functions or procedures for time-to-event analysis. However, both R and SAS include a version of the Mantel-Haenszel test for exact data that is usually called the log-rank test. If it is applied to standard germination data, the log-rank test will give the same results as the Mantel-Haenszel test if there are no seed losses. (To see this, compare equation (S.8) below with equation (S.6), equating the various n and d values in equation (S.8) below to the corresponding N and D values in equation (S.6) and setting weight function $W(t'_i) = 1$ for all i; the variance-covariance matrices are not shown but they, too, are identical in this case.) Moreover, values of the test statistics usually will remain very similar if a small proportion of seeds is randomly lost (roughly 5 % or less, say).

Comparing groups using exact data

Here we provide some additional details about methods for exact data that were not included in the text. Our presentation is based on the lucid account provided by Klein and Moeschberger (2003).

Suppose there are $K \geq 2$ groups, and let $t_1 < t_2 < t_3 < \cdots < t_D$ be the D distinct event times when data from all groups are combined. Let d_{ij} be the number of events at time t_i in group j, let n_{ij} be the number of seeds at risk in group j immediately prior to time t_i, and let $d_i = \sum_j d_{ij}$ and $n_i = \sum_j n_{ij}$. The most commonly used nonparametric tests for comparing groups in time-to-event analysis are based on statistics $Z_j(t')$ given by

$$Z_j(t') = \sum_{i=1}^{D} w(t_i)\left[d_{ij} - n_{ij}(d_i/n_i)\right], \tag{S.8}$$

where $j = 1, 2, 3, \ldots, K$ are the K groups being compared, $w(t_i)$ is a weight function, and t' is the largest time at which $n_i > 0$ for all K groups (so that d_i/n_i is defined for every group). As with the Mantel-Haenszel test discussed above, the total number d_i of germination events at time t_i and the number of seeds at risk in each group j are regarded as fixed for each event time. The term in brackets represents the difference between the observed and expected number of events at the i-th event time in the j-th group under the null hypothesis that the groups do not differ at any of the event times. These terms are then weighted by $w(t_i)$ and summed over all event times.

The actual test statistic employed is constructed from any $K - 1$ of the $Z_j(t')$ (the full set sums to zero and therefore is linearly dependent) and is given by quadratic form Q defined by

$$Q = \mathbf{z}^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \mathbf{z}, \tag{S.9}$$

where $\mathbf{z}$ is the column vector $[Z_j(t')]$ of length $K - 1$, $\mathbf{z}^{\mathrm{T}}$ is its transpose, and $\boldsymbol{\Sigma}^{-1}$ is the inverse of the estimated variance-covariance matrix of $\mathbf{z}$. Under the null hypothesis, $\mathbf{z}$ will have an approximately multivariate normal distribution for large samples, and the distribution of Q will be approximately chi-squared with $K - 1$ degrees of freedom (for additional details, see Klein and Moeschberger, 2003). If all the $Z_j(t')$ values are sufficiently small (i.e., observed and expected numbers of events are very similar at all event times), the value of Q will be close to

zero and the null hypothesis of no difference will be accepted. But if any $Z_j(t')$ is sufficiently large, the value of Q also will be large and the null hypothesis will be rejected.

The various tests based on quadratic form Q differ only in the choice of weight function $w(t_i)$. The most-common tests and associated weight functions are listed in Table 1 of the text.

Effects of random seed loss

Seeds loss includes actual loss of seeds as well as intentional removal due to accidental damage while handling, mold growth, and so on. To assess the effect of such occurrences on nonparametric estimates of the survivor function, we artificially created seed losses in the Japanese knotweed test data as follows. Using an R script, we randomly chose either 5 or 10 of the 100 seeds from each of the three study sites on collection date 3. For each chosen seed, we determined the reported event or censoring time, then randomly chose an earlier time to assign as the loss time. The randomly created artificial loss time for each chosen seed was then substituted for the reported event or censoring time in the data and the survivor function was estimated. Figure S.1 shows representative examples for the Kaplan-Meier estimate. Examples like these suggest that random loss of 5 % or less of seeds usually has only a minor effect on the estimated survivor function. As the percentage of lost seeds increases, it becomes more likely that they will have meaningful effects on the shape or location of the estimated survivor function, as is arguably the case for study site 3 in Figure S.1 with 10 % seed loss.

As noted in the text, another effect of seed loss is that the estimated Kaplan-Meier and life-table survivor functions will no longer be identical at the observation times, because losses are handled differently by the two methods. However, examples with artificially created seed losses suggest that the difference usually will be slight. For example, life-table estimates of survivor functions for the data used to create Figure S.1 differ so slightly from the Kaplan-Meier estimates at the observation times that they are not visually different when plotted on the same panel.

Numerical results suggest that small amounts of random seed loss also have little effect on tests for group differences. For example, Table S.1 shows results of group comparisons for the same data used to create Figure S.1. Note that loss of 5 seeds from the initial total of 100 had a negligible effect on p values, while loss of 10 seeds increased p values by roughly a factor of 3 for 2 of the 3 pair-wise comparisons. Based on examples such as these, a small percentage (roughly 5 % or less) of randomly lost seeds would not be expected to alter the results of statistical tests for group differences unless the evidence for or against group differences was equivocal to begin with.

A graphical test of the proportional hazards assumption

A common graphical method for checking the PH assumption is to plot $-\log(-\log(S(t \mid \mathbf{x})))$ versus t or $\log(t)$ for different values of the covariate vector $\mathbf{x}$ (restricted to values of t such that $0 < S(t) < 1$ so the logarithms remain finite). This method is based on the survivor function, which has a simple form under the PH assumption. Specifically, using text equation (11) in text equation (4), we find that under the PH assumption,

$$S(t \mid \mathbf{x}) = e^{-\int_0^t h_0(\tau)d\tau\ \psi(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})} = S_0(t)^{\psi(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})}, \tag{S.10}$$

Table S.1. Results of group (site) comparisons for seeds collected on date 3 with artificial losses of 0, 5, and 10 seeds out of 100 initial seeds. Results are shown for the two tests from Table 1 of the text that produced the smallest and largest *p* values. Tests were performed with SAS procedure `lifetest`. The three *p* values for each pairwise test are Holm-adjusted for multiple comparisons.

Number of seeds lost	Test	All 3 sites			Sites 1 & 2			Sites 2 & 3			Sites 1 & 3		
		χ^2	df	p	χ^2	df	p	χ^2	df	p	χ^2	df	p
0	Log-rank	41.3592	2	<0.0001	13.2885	1	0.0005	42.4257	1	<0.0001	8.5253	1	0.0035
	Modified Peto-Peto	37.0556	2	<0.0001	13.1640	1	0.0006	38.5451	1	<0.0001	6.1864	1	0.0129
5	Log-rank	40.3026	2	<0.0001	12.9294	1	0.0006	41.3682	1	<0.0001	8.3128	1	0.0039
	Modified Peto-Peto	36.2021	2	<0.0001	12.8097	1	0.0007	37.6639	1	<0.0001	6.1038	1	0.0135
10	Log-rank	31.0984	2	<0.0001	10.2454	1	0.0027	32.0542	1	<0.0001	6.0556	1	0.0139
	Modified Peto-Peto	27.6700	2	<0.0001	10.2201	1	0.0028	28.8992	1	<0.0001	4.1815	1	0.0409

where $S_0(t)$ is the baseline survivor function given by

$$S_0(t) = \exp\left(-\int_0^t h_0(\tau)\, d\,\tau\right). \tag{S.11}$$

It follows from equation (S.10) that

$$-\log(-\log(S(t \mid \mathbf{x}))) = -\log(\psi(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})) - \log(-\log(S_0(t))). \tag{S.12}$$

This function has two terms, the first involving covariates but not time, and the second involving time but not covariates. Therefore, under the PH assumption, different choices of the covariates will produce a family of curves that, when plotted against time, have different elevations (determined by the first term) but the same shape (determined by the second term).

Applying the Cox model to Japanese knotweed data

Exploratory analysis of the data

Plotting nonparametric estimates of survivor functions for different treatment groups is a useful exploratory technique. The examples shown in Figures 4 and 5 of the text clearly suggest that both study site and collection date may have meaningful effects on the temporal pattern of germination.

Checking the proportional-hazards assumption

As noted above, a useful method for assessing the PH assumption graphically is to plot $-\log(-\log(S(t \mid \mathbf{x})))$ versus t or $\log(t)$ for different values of the covariates $\mathbf{x}$ (restricted to values of t such that $0 < S(t) < 1$), using the life-table or Kaplan-Meier estimate of survivor function $S(t)$. The basis for this test is equation (S.12). Roughly parallel curves for different values of the covariates indicate that the PH assumption is plausible, while curves that cross decisively (i.e., two curves change from being clearly distinct with one ordering to being clearly distinct with the reverse ordering) indicate the opposite. Since this method is based on comparison of curves for discrete groups, it requires categorizing quantitative covariates.

Study site is a categorical variable with three categories, while collection date is a quantitative variable with eight distinct values. To obtain discrete collection-date groups for comparison, we grouped the collection dates into two meaningful categories: 1–4 (early collection dates) and 5–8 (late collection dates). Plots of $-\log(-\log(S(t \mid \mathbf{z})))$ versus $\log(t)$ are shown in Figure 6 of the text. (These plots were created by transforming the Kaplan-Meier survivor function produced by R function `survfit()`.) The left panel shows curves for the three study sites, while the right panel shows curves for the early and late collection dates. In both cases, the spacing between the level parts of the step-functions remains similar over time, and we see no evidence of decisive crossing that clearly reverses their ordering. (The estimated survivor functions for sites 1 and 3 cross, but they are so similar that there is never any clear evidence that they actually differ, and hence no evidence of decisive crossing.) We conclude that the PH assumption is plausible for all covariates.

Checking for multicollinearity

We assessed multicollinearity for the Japanese knotweed data by performing a regression analysis using R function `glm()` from the `stats` package with germination time as the dependent variable and site and collection date as covariates, then using the `vif()` function from the `HH` package to estimate the variance inflation factor (VIF) for each covariate. The full set of covariates consists of the original study-site variable (which is categorical) and collection date (which is quantitative). Since there are only two covariates, a single assessment of multicollinearity suffices for the entire model-building process.

The VIF for covariate i, denoted VIF_i, is given by

$$\mathrm{VIF}_i = \frac{1}{1-R_i^2}, \tag{S.13}$$

where R_i^2 is the R^2 value for the linear regression of covariate i on the remaining covariates. Thus, the greater the proportion of variation in covariate i that can be accounted for by the remaining covariates, the closer R_i^2 will be to 1 and the larger (or more inflated) VIF_i will be. Note that multicollinearity only involves relationships among the covariates, so the relationship estimated by `glm()` between the dependent variable and the covariates is irrelevant; `glm()` is used simply to create an object from which `vif()` can determine what proportion of the variation in each covariate is explained by the other covariates (i.e., R_i^2), and from this to determine VIF_i.

A VIF greater than 5 is usually considered to be evidence of multicollinearity, and a value greater than 10 is considered to indicate serious multicollinearity. When we check multicollinearity between the site and collection date variables, we find that VIF = 1 for both covariates, so there is no evidence of multicollinearity.

Testing covariates individually

We now insert the study site and collection date variables individually into the Cox model (using R function `coxph()`) and assess statistical significance (the required p values are reported by `coxph()`). Since study site has three categories, we define two indicator variables, x_1 and x_2, as follows:

$$x_1 = \begin{cases} 1, & \text{if site} = \text{Friends} \\ 0, & \text{otherwise.} \end{cases} \qquad x_2 = \begin{cases} 1, & \text{if site} = \text{Rising Sun} \\ 0, & \text{otherwise.} \end{cases} \tag{S.14}$$

(Note that $x_1 = x_2 = 0$ implies that site = Carroll.) Inserting x_1 (Friends) as a covariate and testing the hypothesis that $\beta_1 = 0$, we find that $p < 0.00001$. Removing x_1 from the model, inserting x_2 (Rising Sun) as a covariate, and testing the hypothesis that $\beta_2 = 0$, we again find that $p < 0.00001$. Therefore, both site variables are candidates for the multivariate Cox model.

The collection date variable is quantitative and therefore can be inserted in the Cox model as a single covariate, x_3. Removing x_2, inserting x_3 as the sole covariate, and testing the hypothesis that $\beta_3 = 0$, we once again find that $p < 0.00001$. Therefore, collection date is a candidate for the multivariate Cox model.

Building a multivariable Cox model

As in the case of multiple regression analysis, various procedures are available for building a

"best" Cox model, and the basis for choosing among them is somewhat subjective. Standard statistical programs such as R and SAS allow one to employ, for example, Akaike's information criterion (AIC), p values, likelihood ratio tests, or combinations of these methods. We illustrate the model-building process with the p-value method, using a forward-selection procedure.

Covariates x_1 (Friends site), x_2 (Rising Sun site), and x_3 (collection date) all were found to have highly statistically significant effects when inserted in the Cox model individually. Next we create a Cox model (continuing to use R function `coxph()`) with the two covariates whose individual p values were smallest; namely, x_3 (smallest p) and x_1 (second smallest p). We test the individual hypotheses that $\beta_1 = 0$ and $\beta_3 = 0$ and in both cases find that $p < 0.00001$. We therefore retain both covariates in the model. Next we add covariate x_2 to the model and test the individual hypotheses that $\beta_i = 0$ for $i = 1, 2, 3$. We find that $p < 0.00001$ for x_1 and x_3 and $p \approx 0.0185$ for x_2. We therefore retain all three covariates. Finally, we sequentially insert the two estimable interaction terms ($x_4 = x_1 \times x_3$ and $x_5 = x_2 \times x_3$), starting with the two covariates with the smallest p values in the 3-variable model. In each case, we test the hypothesis that $\beta_i = 0$ and find that $p > 0.1$, and we therefore exclude both interaction terms. The final Cox model therefore includes all three covariates but no pair-wise interactions. Table 1 of the text summarizes the model, including the estimated values of coefficients β_1, β_2, and β_3.

Note that assessment of multicollinearity above was done using the original categorical covariate for study site rather than the multiple indicator variables by which it is represented in the Cox model. This procedure ensures that the multiple indicator variables representing the original covariate will be included or excluded as a group, as they should be.

Interpreting the final model

The hazard function in the final Cox model has the following form:

$$h(t \mid \mathbf{x}) = h_0(t)\, \exp(-1.276\, x_1 + 0.144\, x_2 + 0.330\, x_3), \tag{S.15}$$

where covariates x_1 and x_2 are indicator variables for study sites Friends and Rising Sun, and x_3 is collection date.

To assess the effect of study site Friends on germination time while controlling for (i.e., removing) effects of collection date, we consider the ratio of the hazard function with $x_1 = 1$, $x_2 = 0$, and x_3 taking on any admissible value (in the numerator) to the hazard function with $x_1 = x_2 = 0$ and x_3 fixed at the same value as in the numerator (in the denominator). The resulting hazard ratio is

$$\mathrm{HR} = \exp(-1.276) = 0.28\,. \tag{S.16}$$

Now, choosing $x_1 = 1$ and $x_2 = 0$ for any x_3 implies that the numerator applies to the Friends site for any chosen collection date, while choosing $x_1 = 0$ and $x_2 = 0$ with the same value of x_3 implies that the denominator applies to the Carroll site for the same collection date. Thus, the hazard ratio in equation (S.16) indicates that for any given collection date, seeds from the Friends site have a hazard function for germination that is 0.28 times as large as (and thus is smaller than) the hazard function for seeds from the Carroll site. This shows that the slope coefficient for Friends implicitly specifies the effect of Friends relative to Carroll. Recalling equation (7) of the text, this result indicates that the survivor function for seeds from Friends will decrease slower with time since sowing than will the survivor function for seeds from Carroll for the same collection date,

and therefore seeds from Friends will tend to germinate later than seeds from Carroll.

The effect of study site Rising Sun with effects of collection date controlled is assessed similarly, except that the values of covariates x_1 and x_2 in the numerator are now $x_1 = 0$ and $x_2 = 1$. The resulting hazard ratio is

$$\text{HR} = \exp(0.144) = 1.15. \tag{S.17}$$

Thus, for any given collection date, seeds from the Rising Sun site have a hazard function that is 1.15 times as large as (and thus is greater than) the hazard function for the Carroll site, showing that the slope coefficient for Rising Sun implicitly specifies the effect of Rising Sun relative to Carroll. It follows that seeds from Rising Sun tend to germinate sooner than seeds from Carroll. In the same way, we can assess the effect of the Rising Sun site relative to that of the Friends site by choosing $x_1 = 0$ and $x_2 = 1$ in the numerator of the hazard ratio while choosing $x_1 = 1$ and $x_2 = 0$ in the denominator. The hazard ratio is then

$$\text{HR} = \frac{\exp(0.144)}{\exp(-1.276)} = \exp(1.420) = 4.14. \tag{S.18}$$

Thus, for any given collection date, seeds from Rising Sun have a hazard function that is $4.14 > 1$ times as large as the hazard function for Friends, and seeds from Rising Sun will therefore tend to germinate sooner than seeds from Friends.

Finally, the effect of collection date with effects of study site controlled can be assessed as follows. We allow site covariates x_1 and x_2 to take on any admissible values but require the values in the numerator and denominator to be the same. Recalling that collection date is a quantitative covariate, we may assess its effect by considering the effect of a unit increment, which corresponds to collecting seeds 2 weeks later. Let $\tilde{x}_3$ denote the value of x_3 in the denominator of the hazard ratio. We allow $\tilde{x}_3$ to take on any individual value in {1, 2, 3, …, 7}, and we require the value of x_3 in the numerator to be 1 greater than the value in the denominator; that is, $\tilde{x}_3 + 1$. Then the hazard ratio is given by

$$\text{HR} = \frac{\exp(0.330\ (\tilde{x}_3 + 1))}{\exp(0.330\ \tilde{x}_3)} = \exp(0.330) = 1.39, \tag{S.19}$$

which indicates that for any given study site, collecting seeds 2 weeks later over the study period increases the hazard function by a factor of 1.39 and therefore tends to reduce germination time.

Should the germination delay be removed?

Scott et al. (1984) assert that the Cox model should not be applied to germination data unless the initial delay is removed from observed germination times, because any difference in delay associated with covariates will prevent in the hazard rates from being exactly proportional. In the Japanese knotweed example, graphical assessment of the PH assumption revealed no clear violation. Nevertheless, visual inspection of the data suggests that the delay in onset of germination may differ slightly among study sites and collection dates. We therefore decided to follow the advice of Scott et al. (1984) and re-run our analyses after removing the initial delay from germination times for the various seed groups.

We removed germination delays as follows. For each of the 24 treatment groups, we determined the shortest observed time to germination among the 100 seeds. In the absence of a delay, the first germination event in every group would have been recorded on the first observation day after the start of the experiment. Therefore, letting $\hat{T}_i$ denote the minimum observed germination time in treatment group i, and recalling that the time between observations was 2 days, we subtracted $\hat{T}_i - 2$ days from each recorded germination time for group i. The minimum germination time in the adjusted data was then 2 days in every group.

We repeated the above model-building procedure using the adjusted data with no delay in germination. The results in every step in the procedure were the same. Thus, all three covariates showed highly significant effects when included in the model individually, with collection date having the smallest p value and the Friends site having the second smallest. Both of these covariates remained highly significant when included together, and all three covariates were significant when Rising Sun was included. Finally, neither of the two estimable interaction terms was significant. The final model, then, includes the same covariates as before and is summarized in Table S.2. Comparing this table with text Table 1 shows that not only are the included covariates the same, but the estimates of coefficients β_i are similar, as well. Clearly, the only thing gained by analyzing the adjusted data is knowledge that our original analysis was robust to any minor differences in germination delay that might exist. This is exactly what is expected, based on the robustness of the Cox model and the fact that graphical assessment of the PH assumption revealed no clear violation.

Table S.2. Summary table of the final Cox model for the Japanese knotweed test data with the germination delay removed. SE denotes the standard error.

Covariate, x_i	Coefficient, β_i	exp(β_i)	SE of β_i	z	p
x_1 (Friends)	−1.203	0.300	0.0784	−15.33	<0.00001
x_2 (Rising Sun)	0.148	1.159	0.0610	2.42	0.01535
x_3 (Collection Date)	0.311	1.365	0.0126	24.68	<0.00001

Including random effects

Applying the modified model-building procedure outlined above to these data, the first step is to include the fixed-effect covariates in a Cox model one at a time, while also including a gamma-distributed frailty term (using the `frailty()` function in R for specifying the frailty term in the model formula of `coxph()`) along with each individual fixed-effect covariate. Thus, each Cox model at this stage of the analysis has both a fixed-effect covariate and a shared-frailty term, where each of the 120 replicates is assumed to be subject to a separate gamma-distributed random effect that applies to all 20 seeds. We find that all three covariates (Friends site, Rising Sun site, and collection date) have significant effects ($p < 0.01$), and that the frailty effect is highly significant ($p < 0.000001$ in all three cases). We therefore include the frailty term in all remaining steps of building the model.

Next we consider the shared-frailty model with the two covariates whose p values were the smallest in the one-covariate models; namely, collection date (smallest p) and Friends site

(second smallest p). We find that both covariates have highly significant effects ($p < 0.000001$ in both cases) and therefore retain both in the model. We then include the third covariate (Rising Sun site) in the model and find that collection date and the Friends site remain highly significant but the Rising Sun site shows no evidence of significance ($p \approx 0.52$). We therefore drop Rising Sun from the model.

Finally, we include the interaction between Friends and collection date in the model. We find weak but not compelling evidence for an effect ($p \approx 0.07$) and therefore drop the interaction from the model.

The final model is summarized in Table 4 of the text. The main difference from the model summarized in text Table 3 is that the covariate representing the Rising Sun site is no longer included in the model. It will be recalled that the germination pattern for this site is roughly similar to that at the Carroll Park site for later collection dates, and that the site slope coefficients in the model implicitly represent effects relative to Carroll Park. Thus, random variation among replicates in the shared-frailty model results in a loss of ability to detect a difference between the Rising Sun and Carroll Park sites. Another way of saying this is that, had the replicates we created for the frailty example been present in the real data, combining data from replicates within treatment groups (and thereby ignoring within-treatment variation that was actually present) would have resulted in detecting a significant difference between the Rising Sun and Carroll Park sites that is not detectable when within-treatment variation due to replicates is accounted for. This additional effect created by ignoring within-treatment variation would be a false positive.

Code examples for nonparametric methods using R and SAS

R code and examples

Data formats

All functions in R package `survival` use the standard data format for modern time-to-event analysis, which assumes the data are exact. Each data record corresponds to one seed and must have at least the following two fields: an event time (either censoring or germination) and a status variable whose value is 0 if the event was a censoring event (lost seed during the experiment, or ungerminated seed remaining at the end of the experiment) or 1 if the event was a germination event. If there are two or more experimental groups or covariates, then additional fields must be included to fully specify these for each seed. An example is shown in Table S.3.

The `lifetab()` function in R package `KMsurv`, which is used to compute the life-table estimate of the survivor function, employs a different input data format that allows the user to ensure germination events are assigned to the proper intervals. Two types of input are required: (1) an array of interval endpoints, the last of which is entered as `NA` ("not available") and represents the "end" of the infinite interval following completion of the experiment, and (2) a table of germination data (with a header row), with each data record (row) including fields specifying any experimental groups used (e.g., study site and collection date, in our test data), the total number of initial seeds in each group, the number of newly germinated seeds found on each observation day, and the number of ungerminated seeds remaining on the last observation day. It is also convenient to include a field specifying the total number of seeds that germinated during

the experiment. If any seeds were lost during the experiment, then each data record also should include the number of seed losses discovered on each observation day. The number of interval endpoints should be one greater than the number of values for germination events.

Table S.3. Data format employed in modern time-to-event analysis. In this example, there are 2 site groups for each of 2 date groups, with 5 seeds per combination. Event times are listed in the "days" column, while the types of corresponding events are indicated in the "status" column: status 1 = germination, status 0 = censuring due to loss during the experiment or failure to germinate by the end of the experiment.

date	site	days	status
1	1	9	1
1	1	13	1
1	1	21	0
1	1	21	0
1	1	21	0
1	2	7	1
1	2	9	1
1	2	11	1
1	2	17	1
1	2	21	0
2	1	5	1
2	1	7	1
2	1	9	1
2	1	9	0
2	1	11	1
2	2	15	1
2	2	5	1
2	2	7	1
2	2	7	1
2	2	9	1

Life-table survivor function

The following R code assumes that germination data are in a CSV (comma-separated values) file named `Germ_data_lifetab.csv`, located in the current working directory, and that there is only one experimental group, no seed losses occurred, the initial number of seeds is in a field named `n.planted`, the cumulative number of seeds that germinated is in a field named `n.germ.total`, and the numbers of germinated seeds found on observation days are in fields 6 through 16.

```
# Load KMsurv library
library(KMsurv)
# Read in germination data in lifetab format
data.df <- read.table("Germ_data_lifetab.csv", header=T, sep=",")
# Create vector of interval endpoints
```

```
t.endpts <- c(0,1,3,5,7,9,11,13,15,17,19,21,NA)
# Create vector of numbers of lost seeds in intervals, with the number for
    the last interval including ungerminated seeds that remain
nlost <-rep(0, length(t.endpts)-1)
nlost[length(nlost)] <- data.df$n.planted - data.df$n.germ.total
# Create vector of numbers of germination events in intervals
nevent <- c(as.vector(data.df[1, 6:16], mode="integer"), 0)
# Create life table
life.table <- lifetab(t.endpts, 100, nlost, nevent)
print(life.table)
# Plot the life-table survivor function
plot(t.endpts[1:12], life.table[, 5], type="o", pch=16, lwd=2, ylim=c(0,1),
    main="Life-Table Survivor Function", xlab="time [days]", ylab="probability
    of not germinating")
# Add approximate point-wise 95% confidence intervals
u95cl <- life.table[, 5] + qnorm(0.025, lower.tail=F)*life.table[, 8]
l95cl <- life.table[, 5] - qnorm(0.025, lower.tail=F)*life.table[, 8]
lines(t.endpts[1:12], pmin(u95cl,1), lwd=1, lty="61", col=gray(0.5))
lines(t.endpts[1:12], pmax(l95cl,0), lwd=1, lty="61", col=gray(0.5))
```

Kaplan-Meier survivor function

The following R code assumes that germination data are in a CSV file named `Germ_data.csv` located in the current working directory, and that the event times and status variables are named `days` and `status` in the file's header row. The Kaplan-Meier survivor function is fitted with function `survfit()`, a plot with 95 % confidence intervals is created in a window on the screen, and a summary is printed to the command window.

```
# Load survival library
library(survival)
# Read in germination data in standard format
data.df <- read.table("Germ_data.csv", header=T, sep=",")
# Fit Kaplan-Meier survivor function
km.fit <- survfit(Surv(days, status) ~ 1, data=data.df, type="kaplan-meier")
# Plot the survivor function
plot(km.fit, lwd=2, col="blue", conf.int=T, xlab="time [days]",
    ylab="probability of not germinating")
# Print summary table
print(summary(km.fit))
```

Comparison of Survivor Functions

The following R code uses function `survdiff()` from the `survival` package to test for homogeneity of a set of three survivor functions and to test for equality of two of the curves, in both cases using the log-rank test. The code assumes that germination data are in the same CSV file used in the previous example and uses data from fields named `days`, `status`, `site`, and `date`, which specify the event times, status variable (germination or loss/censoring), study site, and collection date.

```
# Load survival library
library(survival)
# Read in germination data in standard format
data.df <- read.table("Germ_data.csv", header=T, sep=",")
# Log-rank tests
```

```
# ... Assess all three sites for heterogeneity
test.123 <- survdiff(Surv(days, status)~site, data=data.df, subset={date==8})
# ... Compare sites 1 & 2 only
test.00.12 <- survdiff(Surv(days, status)~site, data=data.df, subset={date==8
    & site!=3})
# Print results on screen
print(test.00.123)
print(test.00.12)
```

SAS code and examples

Data Input

Germination data for time-to-event analysis must be in the standard format described above for R functions in the `survival` package. In the following code example, `proc import` is used to import data from a CSV file named `workdata.csv`.

```
proc import out= work.workdata datafile= "workdata.csv" replace;
getnames=yes;
mixed=no;
scantext=yes;
usedate=yes;
scantime=yes;
run;
```

Life-table survivor functions and comparisons

As noted in the text, SAS procedure `lifetest` allows one to estimate life-table survivor functions, but it does so incorrectly for interval data generated by periodic simultaneous observation unless the data or observation times are adjusted. When estimating the life-table survivor function, SAS assumes the event data are exact events times, which are to be artificially grouped by arbitrary interval limits that are entered separately. Since the only available estimates of event times are the observation times, one is naturally inclined to enter them as the event times. And since these times are in fact the observation times, one is naturally inclined also to enter them again as observation times. This procedure, however, will result in erroneous assignment of germination events to the intervals *following* the correct intervals. It is therefore necessary either to subtract a small (compared to the time between observations) amount from each observation time when entering it as an event time or to add a small amount to each observation time when entering it separately as an observation time. Either procedure forces the values entered as event times to be slightly less than the values entered as observation times, thus assigning events to the correct intervals.

The following SAS code uses the `lifetest` procedure to estimate life-table survivor functions for all study sites, using the `lt` method. Survivor-function plots are requested with the `plots` option. The failure time variable is `days`, the status variable (germination event or loss/censoring) is `status`, and the stratification variable is `site`. The `strata` function is used to compare survivor functions via the log-rank test. SAS performs comparisons using methods that assume exact data, even though the life-table survivor curves are requested. `ods graphics on/off` gives display curves in HTML format. The `freq` statement is used for life-table analysis to determine the total number of failures that occurred in a particular interval. The `intervals`

option specifies endpoints of the observation intervals, which are adjusted here by adding 0.01 to the actual endpoints to ensure events are assigned to the proper intervals. The `outsurv` option creates an ouput SAS dataset called `onea`.

```
ods graphics on;
proc lifetest data=one  method=lt intervals=(5.01 to 21.01 by 2) outsurv=onea
    plots=(s);
time days*status(0);
strata site/test=(all);
freq freq;
run;
ods graphics off;
```

Kaplan-Meier survivor functions and comparisons

The following SAS code uses the `lifetest` procedure to estimate Kaplan-Meier survivor functions for all study sites, using the `km` method. Other options, variables, and functions are explained in the previous example.

```
ods graphics on;
proc lifetest data=one method=km plots=(s);
time days*status(0);
strata site/test=(all);
run;
ods graphics off;
```

Code examples for semiparametric methods using R and SAS

R code and examples

Data format

All functions in R package `survival` use the standard data format for modern time-to-event analysis, which assumes the data are exact. Each data record corresponds to one seed and must have at least the following two fields: an event time (either censoring or germination) and a status variable whose value is 0 if the event was a censoring event (lost seed during the experiment, or ungerminated seed remaining at the end of the experiment), or 1 if the event was a germination event. If there are two or more experimental groups or covariates, then additional fields must be included to fully specify these for each seed. If the data contain replicates, then each replicate in the entire data set must be specified uniquely.

Checking the PH Assumption

The code below checks the PH assumption for categorical variable `site` by first estimating Kaplan-Meier survivor functions with the `survfit()` function, then plotting –log(-log($S(t)$)) versus log(t) for the three survivor functions by transforming the Kaplan-Meier estimates with user-defined function `mlogmlog()`.

```
# Load survival library
library(survival)
# Read in data
data.df <- read.table("Germ_data.csv", header=T, sep=",")
# Define function mlogmlog() to calculate -log(-log(S(t)))
mlogmlog <- function(y){-log(-log(y))}
# Estimate Kaplan-Meier survivor functions for each of the three sites, for
    all collection dates combined
fit.site <- survfit(Surv(days,status) ~ site, type="kaplan-meier",
    data=data.df)
# Plot -log(-log(S(t))) versus log(t)
plot(fit.site, mark.time=F, fun=mlogmlog, log="x", xlab="t [days]", ylab="-
    log(-log(S(t)))", lty=c("solid","longdash","dotted"), lwd=1.75)
```

Checking for multicollinearity

The following code checks for multicollinearity (or collinearity, in this case) between variables `site` and `date`, based on the VIF diagnostic.

```
# Load HH library
library(HH)
# Read in data
data.df <- read.table("Germ_data.csv", header=T, sep=",")
# Fit a generalized linear model predicting days from site and date
multicollinearitycheck <- glm(days ~ site + date, data=data.df)
# Check for multicollinearity among covariates
vif  <- vif(multicollinearitycheck)
# Print the results on the screen
print(vif)
```

Adding covariates to the Cox model

The following code tests for effects of indicator variable `friends` and quantitative variable `date` in a two-variable Cox model, using the `coxph()` function. It then adds an interaction term and tests for its effect.

```
# Load survival library
library(survival)
# Read in data
data.df <- read.table("Germ_data.csv", header=T, sep=",")
# Create a Cox model with two covariates
cox.date.friends <- coxph(Surv(days, status) ~ date + friends, data=data.df)
print(summary(cox.date.friends))
# Now add an interaction term
cox.date.friends.datefriends <- coxph(Surv(days, status) ~ date + friends +
    date:friends, data=data.df)
print(summary(cox.date.friends.datefriends ))
```

SAS code and examples

Data format and input

Germination data for time-to-event analysis must be in the standard format described above for

R. In the following code example, `proc import` is used to import data from a CSV file named `workdata.csv`.

```
proc import out= work.workdata datafile= "workdata.csv" replace;
getnames=yes;
mixed=no;
scantext=yes;
usedate=yes;
scantime=yes;
run;
```

Checking the PH Assumption

The following code checks the PH assumption for categorical variable `site` in the Cox model, using `proc lifetest`.

```
proc lifetest data=interndata1 method=KM plots=(LLS);
time days*status(0);
strata site;
title "Log-log survivor functions";
run;
```

Checking for multicollinearity

The following code checks for collinearity between variables `site` and `date`, using `proc reg` and the VIF diagnostic.

```
proc reg data = interndata1;
model days = site date / vif;
title "Multicollinearity check";
run;
```

Adding covariates to the Cox model

The following code tests for effects of the `site` and `date` variables, using `proc phreg`. It then adds an interaction term and tests for its effect.

```
/*Site and date*/
proc phreg data=interndata1;
model days*status(0)=site date;
run;
/*Interaction*/
proc phreg data=interndata1;
model days*status(0)=site date sitedate;
sitedate=site*date;
run;
```

References

Elandt-Johnson, R.C. and Johnson, N.L. (1980) *Survival models and data analysis*. New York, John Wiley and Sons.

Johnson, N.L., Kotz, S. and Balakrishnan, N. (1997) *Discrete multivariate distributions*. New

York, John Wiley and Sons.
Mantel, N. and Haenszel, W. (1959) Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute* **22**, 719–748.
Scott, S.J., Jones, R.A. and Williams, W.A. (1984) Review of data analysis for seed germination. *Crop Science* **24**, 1192–1199.