Response to Reviewer 3

Thomas F. Whale, Trystan Surawy-Stepaney, and Sarah L. Barr

Colour key. Reviewer comments are shown in blue. Author responses are shown in black. Text quoted from, or explicitly added to, the revised manuscript is shown in green.

General response

We thank all three reviewers for their careful and constructive comments. We agree that the original manuscript presented the log-rank test too broadly and did not sufficiently assess the performance of the Kaplan–Meier confidence intervals or the effects of droplet number and curve shape.

We have therefore substantially revised the manuscript and changed its title to "Survival-analysis methods for droplet-freezing data: Kaplan–Meier confidence intervals and non-parametric curve comparison" to reflect its broader scope. The revised paper assesses the empirical coverage of Kaplan–Meier pointwise confidence intervals and evaluates the false-positive rates and power of several non-parametric comparison tests using realistic fitted freezing curves. We now treat the log-rank test as useful for particular proportional-hazards-like alternatives rather than as a default method. For more general curve-shape differences, we compare established empirical-distribution tests and introduce an Integrated Squared Difference permutation test operating directly on fraction-frozen curves. We have also added Supplementary Information describing the fitted distributions, statistical methods and Python implementation.

We believe these changes provide a more careful and practically useful framework for quantifying uncertainty and comparing droplet-freezing curves, and we are grateful to the reviewers for prompting this substantial improvement.

Detailed response

We thank the reviewer for this careful and constructive review, and for recognising the value of treating droplet-freezing measurements as survival data. We agree with the main methodological concerns raised. In particular, the first manuscript relied too heavily on a small number of illustrative Gaussian simulations, did not test the empirical coverage of the Kaplan–Meier confidence intervals, and treated the ordinary log-rank test more generally than was justified.

We have substantially reworked the manuscript in response. The revised paper now uses fitted representations of three literature droplet-freezing curves spanning steep, intermediate and broad

freezing behaviour. We use repeated simulations to assess the empirical coverage of the Kaplan–Meier intervals, the false-positive rates of the comparison tests, and their power and false-negative rates under shifted, strongly non-proportional and multi-population alternatives. We have also added several non-parametric alternatives to the log-rank test and introduced an Integrated Squared Difference (ISD) permutation test that operates directly on fraction-frozen curves and does not require proportional hazards.

We think these changes address the principal concerns raised by the reviewer and have made the manuscript much stronger.

Major comments

1. Proportional hazards

Reviewer comment

The log-rank test really only delivers full power when hazards are proportional, i.e., when the ratio between two samples is constant in . The manuscript notes this but only flags the case where curves literally cross. In practice the hazard ratio can vary smoothly with even when the curves don't cross, for example, when two chemically different nucleators share the same but have different . PH is really only plausible when the two samples differ in the $n_s(T)$ prefactor and share the same hazard shape, which is rarely strictly true. It would be worth saying so.

Response

We agree. The original manuscript focused too much on literal curve crossing as the main example of non-proportional hazards. As the reviewer points out, proportional hazards may fail whenever the ratio of conditional freezing probabilities changes with temperature, regardless of whether the plotted curves visibly cross.

We have added a substantially expanded discussion of this issue. The revised manuscript now explains that non-proportional behaviour may arise when samples contain multiple ice-active populations, when different populations dominate different parts of the freezing spectrum, when a treatment affects only part of a curve, or when a sample approaches the instrumental background. We also explain why this reduces the power of the ordinary log-rank test.

We have also added simulations that demonstrate this directly. The log-rank test performs well for some relatively simple shifted-curve alternatives in Figs. 5–6, but has much lower power for the steep-versus-broad comparison in Figs. 7–8 and the two-population mixture example in Figs. 9–10.

The revised manuscript therefore no longer recommends the ordinary log-rank test as a general-purpose comparison for droplet-freezing data. We retain it as a useful test when a broadly proportional-hazards-like difference is a reasonable expectation, particularly for relatively simple shifts in freezing behaviour. For more general curve-shape differences, we compare several empirical-distribution tests and introduce the ISD permutation test, which does not require proportional hazards. These changes are reflected throughout the revised statistical background, results, abstract and conclusions.

2. Coverage check

Reviewer comment

The recommendation to report KM bands would be a lot more compelling with an empirical coverage study. If the answer is consistently near 95% the recommendation is on solid ground; if it isn't in some regime, that is itself important to know.

Response

We agree, and this is now a major part of the revised paper. We also now consistently call these pointwise confidence intervals, rather than confidence bands, in response to a comment from another reviewer

We fitted smooth two-component Gaussian mixture distributions to three published droplet-freezing curves: a steep Betula pendula pollen washing water curve, a broader background curve, and a still broader BCS375 feldspar curve. The fitting is briefly discussed int he SI. We then generated repeated synthetic experiments from these fitted distributions. Because the underlying generating curve is known, we can calculate directly how often nominal 95 % Kaplan–Meier intervals contain the value they are intended to estimate.

The results are shown in Figs. 1–3. Coverage is close to the nominal level over most of the fraction-frozen curve for experimentally realistic droplet numbers, but it is not uniform. Coverage deteriorates near the ends of the curve, where the true fraction frozen approaches 0 or 1 and finite experiments contain little information about the exact tail behaviour. This issue becomes more obvious when the intervals are transformed to logarithmic cumulative activity spectra.

We think this provides a much firmer basis for recommending the intervals, while also making their limitations clear. The revised manuscript now states explicitly that they are pointwise confidence intervals, not simultaneous confidence bands, and that the tails of finite droplet-freezing experiments should be treated cautiously.

3. Type I error and power for the log-rank test

Reviewer comment

Figure 4(b) shows one example of a false positive and labels it "Type I error from sampling variability." A single example doesn't tell anyone what the real false-positive rate is.

Response

We agree. The original individual examples were not sufficient to establish either the false-positive rate or the power of the test, and we have removed that approach.

The revised manuscript now estimates false-positive rates using repeated null simulations. For each of the three fitted freezing curves and each droplet number, we generate 1000 pairs of experiments from the same underlying distribution and apply each test at . Figure 4 shows the resulting empirical false-positive rates.

Most tests are close to the nominal 5 % false-positive rate for moderate and large droplet numbers. The clearest deviations occur at very small droplet numbers: the ordinary log-rank test is somewhat liberal at , while the Kolmogorov–Smirnov test is strongly conservative. This calibration exercise is important because it means that differences between tests in the subsequent alternative scenarios can mainly be interpreted as differences in power, rather than simply reflecting one test having a much higher false-positive rate.

We then estimate power and false-negative rates using repeated simulations under realistic shifted and non-proportional situations. This replaces the original reliance on isolated examples of Type I and Type II errors.

4. Only Gaussian simulations

Reviewer comment

Table 1 shows that all the simulated datasets are Gaussian. Real fraction-frozen curves are routinely not and are often long-tailed at the warm end and sometimes bimodal. The "~50 droplets" heuristic in the conclusions should either be tested against some non-Gaussian shapes or qualified as distribution-dependent.

Response

We agree. The original single-Gaussian simulations were too limited to support a general recommendation about droplet number.

The revised simulations use two-component Gaussian mixtures fitted to three literature droplet-freezing curves. The fitted curves span a useful range of realistic behaviour, including a steep freezing transition, a broader background-like curve, and a broad BCS375 curve with freezing spread over a much wider temperature interval. The dual-Gaussian functions are used as flexible empirical representations of realistic curve shapes, not as mechanistic models of ice nucleation.

We have also added a deliberately constructed multi-population scenario in which two curves contain the same two Gaussian freezing components but in different proportions. This produces a shape difference rather than a simple temperature shift and provides a controlled example of non-proportional behaviour.

The revised results show that detectability depends strongly on curve shape. A modest shift in a steep curve can be detected with relatively few droplets, whereas the same shift in a broad curve may remain difficult to detect even with substantially larger droplet populations. We have therefore removed the approximately 50-droplet heuristic and now state that there is no universal droplet-number threshold. Useful precision and statistical power depend on droplet number, curve shape, effect size, temperature region and the test being applied.

5. Quantitative comparison with Fahy et al.

Reviewer comment

Figure 1 shows that the KM bands look similar to the studentized bootstrap. That's good as far as it goes, but if the authors are recommending KM in preference to the existing method then "looks similar" isn't quite enough. A quantitative comparison, say band width at matched temperatures or coverage against a known ground truth, would be straightforward to add and would fall out of the coverage study above.

Response

We agree. The qualitative comparison with the Fahy et al. intervals did not provide a sufficient basis for assessing the statistical performance of the Kaplan–Meier intervals.

Rather than comparing interval widths for a single experimental dataset, we replaced this part of the analysis with the empirical coverage study suggested by the reviewer. Other methods for calculating confidence intervals necessarily involve choices concerning binning, smoothing, resampling, or parameterisation, making a direct comparison dependent on the particular implementation. We therefore considered it more useful to assess whether the Kaplan–Meier intervals themselves perform reasonably. To do this, we generated repeated synthetic experiments from fitted underlying freezing-temperature distributions and calculated directly how often nominal 95 % Kaplan–Meier intervals contained the known underlying value.

We think coverage is the more relevant quantitative test because two methods may produce intervals of similar width while having different coverage properties. Conversely, a wider interval is not necessarily worse if it gives correctly calibrated coverage. The revised analysis therefore focuses on whether the intervals contain the value they are intended to estimate, how that behaviour varies with droplet number and curve shape, and where along the fraction-frozen curve coverage deteriorates.

We have also softened the comparison with previous approaches. The revised manuscript does not claim that Kaplan–Meier intervals are universally preferable to the studentized bootstrap. Instead, we present them as a simple analytical approach that requires no resampling, binning or assumed parametric freezing distribution, and whose empirical pointwise coverage is shown to be sensible over most of the experimentally relevant curve.

6. The patch

Reviewer comment

Carrying the previous interval forward at the last point is pragmatic but it's an arbitrary choice and warrants either some justification or at least a sensitivity check. Does the choice change any of the conclusions on the ash or birch-pollen data? If not, please say so.

Response

Response: We agree that carrying the penultimate interval forward is a pragmatic plotting convention rather than a formally derived confidence interval. We have now made this explicit in the manuscript by adding the following text:

"For plotting purposes, we assign the final point the same interval as the penultimate point. This is a pragmatic plotting convention rather than a formally derived confidence interval at $f(T) = 1$.

It affects only the final plotted point and does not alter the curve-comparison tests or any of the conclusions, because these tests are calculated from the observed freezing temperatures and empirical fraction-frozen curves rather than from the confidence intervals. The final point should therefore not be interpreted as having a formally calibrated confidence interval

The choice therefore does not affect the simulation results or the conclusions of the revised manuscript. It also does not provide additional information for derived cumulative spectra, because quantities based on $-\ln[1 - f(T)]$ are themselves undefined at $f(T) = 1$.

We have removed the volcanic-ash and *Betula pendula* experimental examples from this part of the paper. On reflection, we realised that these individual datasets were not particularly helpful for explaining or validating the behaviour of the method. We have replaced them with the systematic empirical coverage assessment using synthetic experiments generated from fitted literature freezing curves.

Closing comment

Reviewer comment

I think this is a useful methodological contribution and well within scope for the journal. The framing is sound and I expect it to get picked up quickly once it's out. The calibration and assumption-testing points above would, in my view, change a good paper into a reference one. I look forward to the revision.

Response

We thank the reviewer for the encouraging assessment and for the specific suggestions. The comments on empirical coverage, false-positive rates, statistical power, realistic curve shapes and proportional hazards were particularly helpful. They led us to replace the limited illustrative simulations with a much broader calibration and power study, and to rethink the role of the ordinary log-rank test in the paper.

The revised manuscript is no longer primarily a recommendation to use Kaplan–Meier intervals and the log-rank test together. It instead presents a broader non-parametric framework for quantifying uncertainty and comparing droplet-freezing curves, including explicit discussion of the assumptions and limitations of each approach. We think the revision is substantially stronger as a result.