Response to Reviewer 1

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

1. Overall assessment and scope of the manuscript

Reviewer comment

In "Survival analysis for droplet-freezing data: Kaplan–Meier confidence intervals and log-rank tests," Whale et al. present an approach to quantify uncertainty in droplet-freezing data and test differences in frozen fraction curves based on the non-parametric Kaplan–Meier survival function estimator. By applying this method, they are able to derive confidence intervals for frozen fraction

curves and cumulative ice nucleation activity spectra. They also adapt the log-rank test to droplet freezing data to hypothesis test whether two frozen fraction curves are identical or not. They then demonstrate both methods on selected literature and simulated datasets. The method of calculating confidence intervals they present is rigorous, easy to use, and addresses a persistent problem of a lack of standardization in ice nucleation statistics. However, I have much more significant concerns surrounding the use of the log-rank test for ice nucleation data. Specifically, the hypothesis being tested by the log-rank test is limited in the context of ice nucleation, where in many samples there may be multiple different populations of ice active species that dominate at different temperatures in the frozen fraction spectrum. The requirement of proportional hazards is also a concern as many frozen fraction curves in literature do not meet this assumption, further limiting the usefulness of this test. Given that it can also only be applied to frozen fraction curves (and so generally cannot be used to compare across instruments or across samples at different concentrations), I am not convinced of the broad applicability that the authors imply when recommending this statistical test be used. When combined with the fact that the method of calculating confidence intervals has been previously published (Kinney et al., 2024), I am not convinced that this manuscript in its current form constitutes a sufficiently novel and significant contribution to the field. I encourage the authors to consider whether refinements mentioned in the manuscript that address the shortcomings of the log-rank test or other approaches to statistical testing in the context of survival analysis might be more appropriate and useful to a wider range of applications. Specific comments and suggestions are below.

Response

We thank the reviewer for this careful and helpful summary of the manuscript, and for recognising the value of the Kaplan–Meier confidence interval approach. We also agree with the reviewer's main concern. In the original manuscript we presented the ordinary log-rank test too strongly, and did not sufficiently delimit the question it answers or the assumptions under which it is likely to be useful. In particular, we agree that many ice-nucleation datasets may contain multiple ice-active populations that dominate in different parts of the freezing spectrum, and that the proportional-hazards assumption may often be doubtful. We also agree that comparison of fraction-frozen curves is not always the most physically meaningful approach, especially when comparing different concentrations, surface areas, masses, droplet volumes, or instruments.

We have therefore revised the manuscript substantially. The revised paper is no longer framed as a recommendation to use the log-rank test as a general-purpose comparison for droplet-freezing data. Instead, we now present the log-rank test as one useful test for a particular kind of alternative: a broadly persistent, proportional-hazards-like difference between two directly comparable droplet-level datasets. We have added a much fuller discussion of the proportional-hazards assumption, including the point that this assumption can fail even when curves do not visibly cross.

In response to the reviewer's suggestion, we have also added and tested alternative approaches. Most importantly, we introduce an Integrated Squared Difference (ISD) permutation test. This test compares the accumulated squared separation between two empirical fraction-frozen curves over a specified temperature interval. It is designed to address the more general question of whether two directly comparable fraction-frozen curves differ somewhere over the analysed range, without requiring proportional hazards and without allowing differences in different parts of the curve to cancel. We

also compare this test with standard empirical-distribution tests, including Kolmogorov–Smirnov, Cramér–von Mises, and Anderson–Darling tests.

We have added simulation-based assessments of the false-positive rates and false-negative rates of these tests. These simulations use fitted representations of literature droplet-freezing curves, allowing us to test the methods against realistic curve shapes while knowing the underlying generating distributions. The results show that the ordinary log-rank test can perform well for simple shifted-curve alternatives, but can have much lower power for non-proportional or mixture-population cases. This directly addresses the concern that the original manuscript implied broader applicability than was justified.

We also agree that the novelty of the manuscript needed clearer framing, given that Kaplan–Meier confidence intervals were previously applied to droplet-freezing data in Kinney et al. (2024). However, that earlier application resulted from the work developed for the present study and did not provide a general explanation or validation of the method. The revised manuscript now assesses the approach systematically: it explains and implements Kaplan–Meier pointwise confidence intervals for fraction-frozen curves and derived cumulative spectra, tests their empirical coverage using realistic fitted freezing curves with known underlying distributions, and shows how their usefulness depends on droplet number, curve shape, and position within the freezing transition. To our knowledge, this evidence for the practical utility and limitations of Kaplan–Meier intervals for droplet-freezing data is not otherwise available in the literature.

Overall, we think the reviewer's criticism was fair, and it has led us to reframe the manuscript substantially. The revised paper now aims to provide a practical non-parametric framework for uncertainty quantification and curve comparison, rather than promoting the ordinary log-rank test as a default method.

2. Droplet number and the asymptotic approximation

Reviewer comment

The calculation of confidence intervals with the Kaplan–Meier estimator uses the asymptotic normal distribution, which requires that the central limit theorem apply. It is not clear from the manuscript how many droplets are required for this assumption to be accurate. In Figure 2, the authors consider droplet freezing experiments with different numbers of droplets and conclude that the number of droplets used is "essentially a matter of taste" (excluding the simulation with 5 droplets). However, in using this statistical approach, the sample size matters to the accuracy of the confidence intervals – I suggest the authors at least discuss this in more detail and give guidelines on a population size required to meet the central limit theorem requirement, or to achieve a defined statistical power.

Response

We agree. The statement that droplet number was "essentially a matter of taste" was too casual and has been removed. The Kaplan–Meier intervals used here are pointwise intervals based on an asymptotic approximation, so their performance must be checked for the droplet numbers and curve shapes relevant to these experiments.

We have added an empirical coverage assessment using synthetic experiments generated from fitted literature freezing curves. Because the underlying generating curves are known, we can calculate directly how often nominal 95 % intervals contain the value they are intended to estimate. Figures 1–3 show that coverage is close to nominal over most of the curves for experimentally realistic droplet numbers, but deteriorates near the ends of the freezing transition and for very small experiments.

We have also removed any suggestion of a universal minimum droplet number. The revised text now states that useful precision depends on droplet number, curve steepness and the temperature region of interest. Experiments with tens of droplets can provide useful intervals over the central part of steep curves, while broad or multi-population curves generally require more droplets. Droplet number is therefore reported as essential contextual information rather than treated as a simple pass/fail threshold.

3. Differential ice-nucleation activity spectra

Reviewer comment

Is it possible to apply this estimator to differential ice nucleation activity spectra, in addition to the cumulative spectra presented here? This would extend its usefulness.

Response

We agree that this could be useful, but it is beyond the scope of the present study. Kaplan–Meier intervals propagate directly to cumulative quantities because these are monotonic transformations of $f(T)$. Differential spectra require an additional differencing, binning, smoothing or fitting step, and their uncertainty depends strongly on those choices. Developing a rigorously calibrated approach for differential spectra would be a useful subject for future work. We have added the following text after introducing the Kaplan–Meier intervals:

"Kaplan–Meier confidence intervals propagate naturally to cumulative spectra because these are direct transformations of $f(T)$. Differential spectra are less straightforward because they require an additional differencing, binning, smoothing, or fitting step, and their interpretation depends on how the freezing-temperature data are represented, as discussed by Vali (2023) in relation to Fahy et al. (2022). For this reason, we focus here on fraction-frozen curves and cumulative spectra, where the transformation from the Kaplan–Meier estimator is direct."

4. Utility and interpretation of global tests

Reviewer comment

The foremost issue to me is that simply testing whether two frozen fraction curves are different has limited utility and could be misleading in some use cases or if not interpreted carefully. Consider the case of Figure 4c and d. The authors state that the log-rank test produces a correct result in distinguishing the two curves in Figure 4d, while in Figure 4c, the test produces a Type II error. This interpretation is correct under the hypothesis of the log-rank test. However, it is immediately clear from both panels that the curves crossing cause each of the two simulated samples to have

a higher ice nucleation activity at different points in the spectrum. In panel d, it would be far more physically meaningful to say that the curve in red has higher ice nucleation activity at high temperatures, and then meets the curve in grey, which then has a higher ice nucleation activity at low temperatures where its slope is large. These conclusions can be drawn from the confidence bands themselves, and could also be drawn from ice nucleation activity spectra. In contrast, applying the log-rank test tells the researcher that there is merely a difference between the samples, a far less useful result. The authors should further justify and discuss in what situations this specific test should be used and is more useful than simply comparing confidence intervals. They should also caution the reader about the limitations of the hypothesis being tested.

Response

We agree. The original manuscript placed too much emphasis on whether a test rejected the global null hypothesis and did not make sufficiently clear that this is only one part of the interpretation. A significant global test does not identify where the curves differ, which curve is more active over a particular temperature range, or what physical mechanism is responsible.

The revised manuscript now states this limitation explicitly in the abstract, statistical background and conclusions. We have also removed the former language describing individual test results simply as "correct" or "incorrect". The revised analysis instead quantifies false-positive rates, power and false-negative rates over repeated simulations.

We nevertheless think there is value in a formal global comparison. Its role is to ask whether the observed separation between two finite-droplet curves is larger than expected from sampling variability alone. This is particularly useful in marginal cases where curves overlap, droplet numbers are modest, or the freezing transition is broad. Any significant result must still be interpreted using the curves, their confidence intervals and, where appropriate, cumulative spectra.

In response to this and related comments, we no longer recommend the ordinary log-rank test generally. We compare several empirical-distribution tests and introduce the ISD permutation test, which accumulates squared separation between fraction-frozen curves and does not rely on proportional hazards. The aim of the revised comparison framework is not to replace interpretation of the curves, but to provide a clearer and more reproducible statistical complement to it.

We now state in the Conclusions:

"The ISD test is a global test: a significant result indicates that two curves differ somewhere over the analysed temperature interval, but does not identify where they differ or why. Those questions still require inspection of the curves and, where appropriate, comparison of physically normalised quantities such as $K(T)$, $n_s(T)$, or $n_m(T)$."

At the end of the Abstract, we now state:

"We note finally that a significant global test indicates that two curves differ somewhere over the analysed temperature interval, but does not identify where they differ or why; this still requires inspection of the curves."

5. Proportional hazards and non-crossing curves

Reviewer comment

The second issue, which the authors acknowledge, is the requirement of proportional hazards in the log-rank test, i.e. that the ratio of freezing probabilities is constant across the experiment. Notably, the curves need not cross for this assumption to be invalid. The authors argue that violation of the proportional-hazards assumption, or specifically crossing of frozen fraction curves, is unlikely in literature. I disagree. Even in the specific use cases of ageing of laboratory samples that the authors advocate for using the log-rank test, many counterexamples exist in literature. Further discussion is needed in the manuscript to clarify how violating the proportional-hazards assumption influences the results of the log-rank test, and ideally the authors should implement a version of the log-rank test that is robust to non-proportional hazards.

Response

We agree. The original manuscript treated visibly crossing curves as the main example of non-proportional hazards and implied that such behaviour was relatively unusual. That was not justified. As the reviewer notes, proportional hazards can fail without literal curve crossing whenever the relative freezing probabilities change across the temperature range—for example, because of ageing effects, background freezing, or multiple ice-active populations.

We have removed the suggestion that non-proportional behaviour is uncommon and added a fuller explanation of how violation of proportional hazards affects the ordinary log-rank test. In particular, signed observed-minus-expected contributions from different parts of the temperature range can partly cancel, substantially reducing power.

We have also added simulations demonstrating this directly. The log-rank test performs well for some 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 mixture-population example in Figs. 9–10. The revised recommendation is therefore to use the log-rank test only when a proportional-hazards-like difference is a reasonable expectation. For more general curve-shape differences, we recommend the ISD test or an appropriate empirical-distribution test.

6. More difficult and multi-population simulations

Reviewer comment

The data used to demonstrate these approaches are generally well-behaved for this test, except for the data in Figure 4c and 4d, which break the assumptions of the test in very controlled, specific ways. I would encourage the authors to test the log-rank test with other, less well-behaved simulated datasets such as those generated from multiple separate Gaussians, and to demonstrate the limits of the log-rank test as a function of population size.

Response

We agree, and have replaced the limited illustrative examples in the original manuscript with a broader simulation study. The revised simulations use fitted literature-like curves spanning steep,

intermediate and broad freezing behaviour, together with deliberately constructed shifted, strongly non-proportional and two-component mixture scenarios.

The mixture example in Figs. 9–10 contains the same two Gaussian freezing populations in different proportions and was designed specifically to test the situation described by the reviewer. We now report false-negative rates as a function of droplet number, showing directly how power depends on sample size, curve shape and the magnitude and form of the difference. We have also added null simulations in Fig. 4 to establish that differences in power are not simply caused by large differences in false-positive behaviour.

7. Fraction-frozen curves versus physically normalised spectra

Reviewer comment

It is not clear to me what the benefit of testing differences between specifically two frozen fraction curves is, as opposed to the volume-, surface-area-, or mass-normalised cumulative ice-nucleation activity spectra $K(T)$, $n_s(T)$, or $n_m(T)$. A stronger argument should be made for the need for using this specific test. Other rigorous approaches to comparing differences between ice-nucleation spectra already exist, and confidence intervals or confidence bands can also be compared. Therefore, a stronger argument for using the log-rank test on frozen fraction curves over these other approaches should be presented in the text.

Response

We agree with much of this comment, and it has helped us clarify what the curve-comparison tests are for. In the original manuscript we did not explain this well enough, and we agree that we implied the log-rank test was more generally useful than it is. We also agree that simply asking whether two fraction-frozen curves are "different" is a limited question. A significant global test does not say where the curves differ, which sample is more active over a particular temperature interval, or what physical mechanism is responsible. Those questions still require inspection of the plotted curves and, where appropriate, comparison of normalised spectra.

However, we do think there is value in quantifying the evidence that two curves differ, provided the result is interpreted carefully. The purpose of a formal test is not to provide the physical interpretation of the difference, but to ask a narrower statistical question: is the observed separation between two finite-droplet curves larger than would be expected from sampling variability alone if the droplets were drawn from the same underlying freezing-temperature distribution? In that sense, a p value gives a reproducible measure of evidence against the no-difference null hypothesis, with a defined false-positive rate. This can be useful in marginal cases where curves overlap, droplet numbers are modest, or the freezing transition is broad.

In the Abstract we now state:

"These tests quantify whether the observed separation between two finite-droplet curves is larger than expected from sampling variability alone, while recognising that a significant global test does not identify where the curves differ or why."

In the Introduction we now state:

"Although droplet-freezing datasets can be analysed in several defensible ways, the field lacks a widely adopted statistical framework for reporting uncertainty and quantitatively comparing fraction-frozen curves. Decisions about whether two curves represent meaningfully different freezing behaviour are often based on visual inspection or informal criteria. Visual comparison of fraction-frozen curves remains essential, because it shows where curves differ and suggests possible physical interpretations. However, visual comparison alone can be ambiguous when droplet numbers are small, curves are broad, or confidence intervals overlap over part of the temperature range. It also does not provide a pre-specified false-positive rate or an estimate of statistical power. Formal tests are therefore useful as a complement to graphical interpretation, indicating whether the observed separation between two finite-droplet curves is larger than expected from sampling variability alone."

We have also changed the paper's position on the log-rank test. We no longer argue that the ordinary log-rank test should be used as a general comparison for fraction-frozen curves, or that it is preferable to comparison of physically normalised spectra. Instead, we present it as useful only when a proportional-hazards-like difference is a reasonable expectation, such as a broadly shift-like change in freezing behaviour. For more general curve-shape differences, we now compare several non-parametric alternatives and introduce the ISD permutation test, which directly tests the accumulated separation between the plotted fraction-frozen curves.

Finally, we agree that normalised spectra are usually the more meaningful quantities when comparing across concentrations, surface areas, masses, droplet volumes, or instruments. We have added this caveat explicitly. The revised manuscript therefore does not imply that global tests on $f(T)$ replace physically normalised spectra.

Specific comments

8. Priority of Monte-Carlo confidence intervals

Reviewer comment

Line 187: These Monte-Carlo confidence intervals were first presented earlier than Whale et al. 2022, for example in the supplemental information of Jahl et al. (2021). I recommend removing this line, as the earliest use of this method is not important to this study.

Response

We agree. This was clumsy wording rather than an intended claim of priority. We were referring to the particular implementation used in the original comparison. That comparison is no longer central to the revised paper and the sentence has been removed. The revised manuscript instead assesses Kaplan–Meier interval performance directly through empirical coverage simulations.

9. Pointwise intervals versus simultaneous confidence bands

Reviewer comment

In the figures of this paper, the Kaplan–Meier confidence intervals are presented as continuous bands,

apparently through interpolations between the calculated confidence intervals. This should be stated somewhere, as continuous confidence bands are not precisely equivalent to interpolated discrete confidence intervals (which are less accurate than true confidence bands; see Sachs et al., 2022).

Response

We agree and thank the reviewer for identifying this. We have revised the manuscript throughout to describe the intervals as pointwise confidence intervals, not confidence bands. The shaded areas shown in the figures are visual connections between pointwise intervals and should not be interpreted as simultaneous confidence bands. This distinction is now stated explicitly, with reference to Sachs et al. (2022).

10. Software and example data

Reviewer comment

The provided script and example data work on my machine, and the methods used are appropriate. I appreciate the authors providing an easy-to-use script for this approach as a supplement to the manuscript.

Response

We thank the reviewer for checking the script and are pleased that it worked as intended. We have updated the accompanying notebooks and scripts to reflect the revised manuscript. The updated material now includes separate notebooks for calculation of Kaplan–Meier pointwise confidence intervals, the ordinary log-rank test, and the new ISD permutation test.