Response to Reviewer 2

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 the careful and constructive review, and for recognising the need for clearer and more rigorous statistical treatment of droplet-freezing measurements. We agree with the two main concerns raised here. First, the original manuscript did not do enough to establish how the Kaplan–Meier pointwise confidence intervals behave as a function of droplet number and curve shape. Second, we presented the ordinary log-rank test too strongly without adequately addressing the proportional-hazards assumption or comparing it with suitable alternatives.

We have revised the manuscript substantially. The revised paper now includes an empirical coverage assessment of the Kaplan–Meier intervals, repeated simulations of false-positive rates and power, fitted multi-component freezing distributions based on literature data, and several non-parametric alternatives to the log-rank test. We have also introduced an Integrated Squared Difference (ISD) permutation test, which compares the accumulated separation between two fraction-frozen curves directly and does not require proportional hazards. The log-rank test is retained as a useful method for particular shift-like alternatives, but is no longer recommended as a general-purpose comparison for droplet-freezing data.

Major comments

1. Population size and the central limit theorem

Reviewer comment

How large does the population have to be for the central limit theorem to hold true, i.e. for which population sizes can you recommend your method of calculating the confidence intervals?

Response

We agree that the original treatment of this issue was inadequate. In particular, the statement that the number of droplets was "essentially a matter of taste" was too casual and has been removed.

There is no sharp droplet-number threshold at which the asymptotic approximation suddenly becomes valid. Its usefulness depends on the number of droplets, the shape and steepness of the underlying freezing curve, and the part of the curve being interpreted. We have therefore addressed this empirically rather than trying to assign a universal central-limit-theorem threshold.

In the revised manuscript, we generate repeated synthetic experiments from fitted representations of three literature freezing curves. Because the underlying generating curves are known, we can calculate directly how often nominal 95 % Kaplan–Meier pointwise confidence intervals contain the values they are intended to estimate. These simulations show that the intervals behave reasonably well over much of the main freezing transition for droplet numbers typical of many experiments, but that coverage deteriorates near the tails, where finite experiments contain little information. Very small experiments produce coarse curves and broad intervals and should be treated cautiously.

We now state that experiments with tens of droplets can provide useful pointwise uncertainty estimates over the main freezing transition, particularly for steep curves, but that broad, long-tailed or multi-population curves generally require more droplets for comparable precision. We deliberately avoid presenting approximately 50 droplets, or any other number, as a universal threshold.

Reviewer comment

What are typical population sizes in the literature for frozen fraction curves?

Response

We agree that this context is useful and have added the following text to the Introduction:

"Many droplet-freezing datasets contain tens to hundreds of droplets per measurement, although the number varies widely with instrument type. Microlitre-scale assays typically freeze on the order of

40–100 droplets per experiment (Whale et al., 2015; Budke and Koop, 2015). High-throughput plate systems can use several hundred wells (Miller et al., 2021; Kunert et al., 2018), and microfluidic systems can generate hundreds to thousands of droplets (Stan et al., 2009; Tarn et al., 2025). On occasion, measurements with fewer than 10 freezing events in larger water volumes are reported, e.g. Kobayashi et al. (2016)."

This wide range reinforces the need to report droplet number alongside fraction-frozen curves and confidence intervals. The apparent smoothness of a plotted curve can otherwise obscure substantial differences in the statistical resolution of different experiments.

2. Applicability and assumptions of the log-rank test

2a. Proportional hazards and crossing curves

Reviewer comment

The log-rank test assumes proportional hazards, which is not given for ice nucleation. In line 270 the authors discuss that a crossing of two frozen fraction curves is unlikely, but that is certainly not true in general. Especially when analysing samples with multiple INP populations and comparing different populations, crossing will occur quite frequently. The authors also propose a way to handle those cases and I think this should be part of the manuscript.

Response

We agree. The original manuscript placed much too much emphasis on literal curve crossing and incorrectly suggested that such behaviour would be unusual. We have removed that statement.

The revised manuscript now explains that crossing is only one obvious visual example of non-proportional hazards. The proportional-hazards assumption may also fail when curves do not cross, for example when different ice-active populations dominate different parts of the freezing spectrum or when a sample approaches the instrumental background at lower temperatures.

We have added simulations specifically addressing these cases. These include a comparison between steep and broad fitted freezing curves and a controlled two-component mixture example in which two samples contain the same freezing populations in different proportions. The results show that the ordinary log-rank test can lose substantial power in such cases.

We have also implemented the broader approach requested by the reviewer. The revised paper compares the log-rank test with Kolmogorov–Smirnov, Cramér–von Mises and Anderson–Darling tests, and introduces an integrated squared fraction-frozen distance (ISD) permutation test. The latter compares the accumulated squared separation between two empirical fraction-frozen curves and does not require proportional hazards. The revised manuscript therefore no longer presents the ordinary log-rank test as a general solution for crossing or multi-population curves.

2b. Non-proportional hazards without curve crossing

Reviewer comment

Even in the case of two frozen fraction curves not crossing, the assumption of proportional hazards is not a given, for example when one population freezes at higher temperatures, while the other population freezes at lower temperatures. Figure 4 compares different results of the log-rank test, but generally focuses on the comparison of curves that have very similar means. One could make a case that the log-rank test could be a good way to check if two frozen fraction curves originate from the same distribution in the case of similar means. Consider adding this in your final statements about the recommendation to use the log-rank test for general comparison.

Response

We agree that the original manuscript did not distinguish clearly enough between visible crossing and the more general proportional-hazards assumption. This has now been corrected throughout.

The revised text explains that proportional hazards can fail whenever the ratio of conditional freezing probabilities changes with temperature, whether or not the plotted curves cross. We therefore do not recommend the log-rank test simply because two curves have similar means, nor do we assume that non-crossing curves necessarily satisfy proportional hazards.

Instead, the ordinary log-rank test is now described as most useful when the expected difference is a broadly persistent, proportional-hazards-like change in freezing probability. A relatively simple temperature shift is one important example. Figures 5–6 show that the log-rank test can perform well for these shifted-curve alternatives. In contrast, Figs. 7–10 show that it can perform poorly for curve-shape and mixture-population differences.

We think this provides a clearer basis for deciding when the log-rank test is likely to be useful. It is retained as one test for a particular type of alternative, rather than recommended as a general comparison for all fraction-frozen curves.

2c. Usefulness of censoring

Reviewer comment

One of the main features of the log-rank test, i.e. the ability to use it for censored data, is discussed in the manuscript, but is to my knowledge not useful for any experimental data of common instruments in ice-nucleation research. Can the authors think of an experimental setup where this property could be useful?

Response

We agree that censoring is unlikely to be an important advantage for most conventional droplet-freezing experiments, and we have reduced the emphasis placed on it.

A possible use would be an experiment that ends before all droplets have frozen, for example because the instrument reaches its minimum temperature. Droplets remaining unfrozen at the end of the experiment could then be treated as censored observations rather than assigned an artificial freezing temperature. Censoring could also arise if individual droplets became unobservable or were lost during an experiment, although such cases would need to be considered carefully because the censoring mechanism should not itself be related to freezing behaviour.

In practice, good experimental design should usually avoid these situations. We now present the ability to accommodate censoring as a technical feature of the Kaplan–Meier framework rather than as a major motivation for its use in ice-nucleation experiments.

3. Comparison with previously used approaches

Reviewer comment

It needs to be shown that the log-rank test actually is better than previously used methods. Of course this is a given for "visual inspection", which is purely subjective and not scientific. For the ad hoc statistical comparisons, the authors do not provide examples. Even just a calculation of the absolute error between the two curves is a statistical approach that can certainly provide a quantitative result, which—combined with additional parameters—offers a more detailed comparison of the two curves. Providing examples and comparing them to their approach would certainly strengthen the manuscript and the case for using the log-rank test.

Response

We agree that the original manuscript did not justify presenting the ordinary log-rank test as preferable to other quantitative approaches, and we no longer make that claim. We also do not think there would be much value in systematically comparing the revised methods with the ad hoc criteria previously used in the literature. These include, for example, comparing a shift in median freezing temperature with a multiple of the instrument temperature uncertainty. Such approaches may be useful for particular experiments, but they do not provide a general test of whether two complete fraction-frozen curves differ.

Instead, the revised paper compares the log-rank test with several established non-parametric tests and the new ISD permutation test. We assess their false-positive rates and power under several realistic curve shapes and alternatives. The results show that no single test is best in every situation, and the revised manuscript recommends choosing a test that is reasonably matched to the scientific question rather than treating the log-rank test as generally superior.

General comments

4. Upright sub- and superscripts

Reviewer comment

Sub- and superscripts that refer to an abbreviation should be upright. For example, as shown in Eq. (7): $N_{U,\text{tot}}(T)$.

Response

We agree and have revised the mathematical notation throughout.

5. Warmer and colder temperatures

Reviewer comment

Temperature can be high or low, but a temperature can never be warm or cold. An example is shown in lines 170–171.

Response

We appreciate the reviewer's preference, but we have retained "warmer" and "colder" where these terms improve clarity. These are established usages in scientific English, and in ice-nucleation work they help avoid ambiguity between numerical temperature and the magnitude of supercooling. For example, "warmer-freezing component" unambiguously describes a component active closer to 0 °C, whereas "higher" or "lower" can be interpreted in several ways for negative temperatures. We have nevertheless checked each use and replaced the terms with "higher temperature" or "lower temperature" where that wording is clearer.

6. Mathematical variables

Reviewer comment

Variables should be italic; this is also true for Greek symbols. An example is shown in line 212.

Response

We agree and have checked the equations and surrounding text for consistent mathematical formatting. Variables and Greek symbols are now italicised, while descriptive labels and abbreviations are set upright.

7. Table formatting

Reviewer comment

Tables should generally be set with the least amount of vertical and horizontal lines. While this is ultimately a stylistic choice, I would recommend removing vertical lines and all horizontal lines apart from the top rule, the middle rule and the bottom rule.

Response

We thank the reviewer for this suggestion. The tables referred to in the original manuscript have been removed as part of the substantially revised analysis. Any remaining tables use a simpler layout without unnecessary vertical or horizontal rules.

8. Number and unit spacing

Reviewer comment

There should be a space between the number and the unit. An example is shown in line 219.

Response

We agree and have checked the presentation of numerical values and units throughout. Following the convention used by Copernicus, a space is placed between the numerical value and the complete unit symbol, for example 2 °C. There is no space between the degree symbol and C.

Specific comments

9. Logarithm at zero

Reviewer comment

Line 120: The authors state that the logarithm cannot be calculated when its argument is zero. While this is true, a more accurate statement would be that the logarithm is not defined in such a case.

Response

We agree and have changed the wording to state that the logarithm is not defined when its argument is zero.

10. Random-number generator website

Reviewer comment

Line 206: There are many methods available to generate random numbers, also for example within the Python framework. Can the authors provide an explanation for the use of the given website?

Response

We agree that there was no good reason to use an external website for this purpose. The website and associated discussion have therefore been removed. The revised simulations use pseudorandom-number generation within Python, rather than an external random-number service.

Comments on tables and figures

11. Table 1 precision

Reviewer comment

Table 1: The amount of significant digits should be equal across the table and especially for the mean and the standard deviation.

Response

We agree. The original table has been removed during the substantial restructuring of the manuscript.

Numerical values reported in the revised tables, figures and captions have been checked for consistent and appropriate precision.

12. Figure 2 presentation and colour accessibility

Reviewer comment

Figure 2: The (c) is shifted slightly more to the left than to the right. The colours should also be chosen with consideration of colour deficiencies. While the use of different markers is good, the two black squares shown in panels (c) and (d) should be of the same size.

Response

We agree. The original figure has been replaced as part of the revised simulation analysis. The new figures use consistent panel positioning, marker sizes and line styles. We have also used combinations of colour, marker shape and line style so that the figures do not rely on colour alone to distinguish datasets or tests.

13. Figure 3 labels, colours and ticks

Reviewer comment

Figure 3: Here the label is inconsistent with Figure 2, where "Fraction frozen" was used. The same axis labels should be chosen throughout the manuscript. The colours should also be chosen with consideration of colour deficiencies. The amount of tick labels should also be reduced, i.e. every 4 K.

Response

We agree. The original figure has been replaced, and the revised figures have been checked for consistent use of "fraction frozen" or "fraction unfrozen", depending on the quantity actually plotted. Axis labels, tick spacing, colour choices and line styles have also been standardised throughout the revised manuscript.

14. Table 2 precision and p values

Reviewer comment

Table 2: Same comment as for Table 1. In addition, I do not see the value of reporting the numerical value of the p value.

Response

We agree that the original table reported more numerical precision than was useful. The table has been removed during the restructuring of the manuscript.

Where p values are reported in the analysis notebooks, they remain available for reproducibility, but the manuscript does not present long numerical p values that imply unjustified precision.

15. Figure 4 resolution

Reviewer comment

Figure 4: The graphic is very blurry and a version should be provided with more pixels.

Response

We agree. The original figure has been replaced by newly generated figures for the revised simulation study. All revised figures have been exported at higher resolution and checked for legibility.

Technical comments

16. Approximate symbol

Reviewer comment

Line 286: The approximate symbol should be used: "..., whereas $\approx$ 50 droplets...".

Response

We agree. This sentence has been removed as part of the revision because we no longer present approximately 50 droplets as a general threshold. Where approximate numerical values remain elsewhere in the manuscript, the appropriate symbol or wording is now used.

17. Missing full stop

Reviewer comment

Line 298: Full stop missing.

Response

Corrected.