We thank all the reviewers for their helpful comments and feedback.

To Reviewer 1 **Relevance of application domains; distinctions between Sec.2 and 3.** Quasi-(in)dependent data, where one event occurs after another, are very common in many fields, e.g., biomedical studies, social sciences, marketing analyses, etc. A test of Quasi-Independence (QI) thus has a broad range of practical applications.

QI and Right-Censoring (RC) are very different data properties. Firstly, QI is a deterministic hard constraint ($X \leq Y$), while RC is a stochastic property of the data (incomplete observations). In this paper we solve the problem of testing QI, giving a general solution in Sec.2. Additionally, since many real-life scenarios display RC (especially biomedical), we extend our approach (given in Sec.2) to handle RC data (Sec.3). We will better emphasise the distinction in the final version.

Why a factorizing kernel? There are two reasons: (1) This choice is sufficient to solve the problem: our resulting $\mathfrak{K}$ is c_0-universal, and Theorem 4.2 shows power goes to 1 asymptotically, meaning our test correctly rejects the null hypothesis for any dependency (with enough data). This is illustrated in our experiments on complex scenarios. (2) The resulting test statistic has a simple expression, Equation (7), which leads to a computationally efficient test.

Using a kernel that directly incorporates censoring? This is an interesting question. Note that a deterministic kernel (such as ours) has no way of encoding random censoring. Thus, in order to allow the kernel to correct the bias due to RC, we would need to consider a random kernel. Random kernels are much harder to analyse, thus we leave this as a potential future research direction.

Bias from censoring and other form of censoring? First a point of clarification: our test-statistic uses *all* the data (X, T, Δ), and not only observations biased by the multiplication by the indicator function, i.e, $(X\mathbf{1}_{\{Y<C\}}, T\mathbf{1}_{\{Y<C\}})$. That said, we would be very interested in extending our work to other forms of censoring, such as interval censoring.

Experiments on real data. We believe that the real data experiments show an exciting result: on the abortion dataset, with a very high level of censoring (90%), our test is the only one to correctly reject the null, and detect the signal in the presence of such heavy censoring. This strong performance under heavy censoring is further confirmed in our appendix G, fig. 8, showing on synthetic data that we can correctly detect quasi-dependence even with high censoring levels.

To Reviewer 2 **Dependency beyond ordering.** Thanks for your suggestions for clarifying the presentation of the QI setting. We will include a more detailed explanation, and some examples illustrating the problem.

To Reviewer 3 **Quantify degree of quasi-dependency instead of testing?** We agree that a measure of the degree of quasi-dependence would be very valuable, and an interesting research direction. There remain settings where a binary decision can be important, however: in particular, QI is associated with simpler models, which can be employed if the test accepts the null, but not if the null is rejected. This is the case for [28], where (unlike prior works) the authors don't ignore quasi-dependence, and obtain completely different results.

Alternative forms of F_X and S_Y can be used? While Equation (1) formally defines quasi-independence, deriving the specific form of $\widetilde{F}_X$ and $\widetilde{S}_Y$ can be complex, and will depend heavily on the specific problem setting; see the proof of Lemma D.2 in the supplement.

Link to the log-rank test. Log-rank tests are best known in the two-sample test setting, however they can be easily extended to other settings (continuous time, covariates, etc), by exploiting their relationship with score tests. This general idea was used by Emura and Wang [9] in obtaining the term inside the parenthesis of Equation (4), known as the weighted log-rank statistic (with weight function ω).

Kernel choice with temporal ordering. The kernel does not need to be designed to incorporate the constraints implied by the temporal ordering, or by RC. All the information about the data is fed through the functions ρ and ρ^c, which impose the relevant constraints arising from order and RC. The kernel defines the space of functions ω used to test the null hypothesis. We use a c_0-universal (Theorem 4.2) kernel, for which the space is rich enough that the power goes to 1 asymptotically for any alternative.

Computational complexity. To compute the kernel based statistics of sample size n, it takes $O(n^3)$ with efficient computation of $\hat{\pi}$ and A, B matrices. A testing procedure with m wildboostraps takes $O(mn^2)$ to compute. Thus, overall cost is $O(n^3 + mn^2)$.

Computational cost comparison of WLR and WLR_SC. The weighted log-rank test can be seen as a special form of the proposed test, with a constant-valued kernel. Computing WLR and WLR_SC takes $O(n^3)$ (same as KQIC). Instead of using wild bootstrap for the test threshold, however, Emura and Wang [9] suggests obtaining the rejection threshold by approximating the relevant integrals empirically, which also costs $O(n^3)$. The overall cost is then $O(n^3)$.

To Reviewer 4 **Open-source code.** We will provide a link to the source code in the final version of the paper.