Accurate Evaluation of Quickest Changepoint Detectors via Non-parametric Survival Analysis

Taiki Miyagawa [1] **Akinori F. Ebihara** [1]

Abstract

We propose non-parametric estimators for the average run length (ARL) and average detection delay (ADD) in quickest changepoint detection (QCD) under finite and irregular sequence lengths. Although ARL and ADD are widely used as optimality criteria in theoretical and simulation studies, their application to real-world datasets is hindered by limited and irregular sequence lengths. To address this issue, we propose non-parametric estimators for the ARL and ADD, termed *KM-ARL and KM-ADD*, by drawing an analogy between QCD and survival analysis to model detection probabilities under sequence truncation. We derive estimation bias bounds and prove that they are asymptotically unbiased unless extrapolation is required. Experiments on simulated and real-world datasets demonstrate their practical utility, enhancing robustness against limited and irregular sequence lengths, improving interpretability, and facilitating empirical, intuitive model selection. Our Python code is provided at https://github.com/TaikiMiyagawa/Kaplan-Meier-Average-Run-Length, offering ready-to-use implementations for practitioners.

1. Introduction

Progress in machine learning depends not only on better models but on reliable performance measurement (Hernàndez-Orallo et al., 2004; Lachiche et al., 2005; 2006; Pardo-Fernández & Castro, 2025). Without principled estimation, improvements may be illusory and decisions may be incorrect. Therefore, reproducible and interpretable evaluation is a first-class concern in machine learning, not an afterthought.

We study evaluation metrics for models in online quickest changepoint detection (QCD) with unknown pre- and post-change distributions, where their datasets with changepoint labels are available. QCD has been extensively studied theoretically, with many models proposed and their optimality proven (Tartakovsky, 2019). It also has diverse real-world applications, including statistical process control (Hawkins et al., 2003), industrial quality control (Wadinger et al., 2024), epidemiology (Johnson & Pedersen, 2025), wireless sensor networks (Hadjiliadis et al., 2009), health monitoring (Tan et al., 2023), radar target detection (Xiang et al., 2021), and seismic sensing (Li et al., 2016).

The average run length (ARL) and the average detection delay (ADD) are central to the theoretical and simulation analysis of QCD models (Tartakovsky et al., 2014; Tartakovsky, 2019). The ARL is defined as the average time a detector takes to raise a false alarm, while the ADD refers to the average delay a detector takes to identify a changepoint after it has occurred. These metrics exhibit a tradeoff: reducing ADD typically shortens ARL (making the detector trigger-happy) and vice versa.

Although the ARL and ADD are widely used as optimality criteria in theoretical and simulation studies, their application to real data is challenging because practical sequences are often limited and irregular in length. Fig. 1(a) shows that naive ARL and ADD estimators yield substantial bias and variance, hindering reliable evaluation of QCD models.

To address this issue, we draw an analogy between QCD and survival analysis (Kleinbaum & Klein, 1996) and propose non-parametric estimators for the ARL and ADD. We adapt the Kaplan-Meier estimator (KME) (Kaplan & Meier, 1958) to the QCD setting to model detection probabilities under sequence truncation. These estimators are non-parametric, requiring no assumptions on the underlying data distribution (e.g., exponential). We derive estimation bias bounds for these estimators, termed *KM-ARL* and *KM-ADD*, by decomposing the estimation bias into *finite-sample bias* and *truncation bias*. We then show that the finite-sample bias decays exponentially with increasing dataset size and that the truncation bias is smaller than that of conventional

[1]NEC Corporation, Japan. Correspondence to: T. Miyagawa <miyagawataik@nec.com>.

Proceedings of the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

estimators. Building on these findings, we prove that the KM-ARL and KM-ADD are *asymptotically unbiased* unless extrapolation is required.

To demonstrate practical applicability, we conduct experiments on both simulated and real-world datasets with limited and irregular lengths. The results show that our estimators reduce estimation bias compared to baseline estimators and enhance robustness to limited and irregular sequence lengths, thereby improving interpretability and facilitating empirical, intuitive model selection (Fig. 1(b)). Our code is provided at `https://github.com/TaikiMiyagawa/Kaplan-Meier-Average-Run-Length`, offering off-the-shelf, ready-to-use implementations in Python (Van Rossum & Drake, 2009) for practitioners.

Our contributions are threefold: (1) we propose KM-ARL and KM-ADD, non-parametric estimators for the ARL and ADD, enabling their evaluation on real-world datasets with limited and irregular-length sequences; (2) we derive estimation bias bounds for these estimators and prove that they are asymptotically unbiased unless extrapolation is required; and (3) we demonstrate their practical utility through experiments and provide ready-to-use Python implementations.

2. Related Work

We highlight our contribution in the context of prior research on ARL and ADD estimation using survival analysis. To our knowledge, in QCD, no metric in its vanilla form explicitly accounts for irregular sequence lengths. Sahki et al. (2020) empirically estimate ARLs and ADDs on truncated sequences of *fixed* length, leveraging *parametric* survival analysis, assuming the survival function decays exponentially. Bradley et al. (2023) employ a *semi-parametric* survival analysis method, the Cox proportional hazards model (Cox, 1972), which assumes an exponential response of the estimator to covariates. They also use the KME to empirically evaluate intrusion detectors under *fixed* length truncation; however, the ARL and ADD are not directly estimated, and no theoretical analysis is provided. Lim & Lee (2025) propose ad hoc estimators for the ARL and ADD on sequences of *fixed* length; however, their theoretical justification has yet to be established, as these estimators diverge when the sequence length $\to \infty$. In contrast, we build our estimators on *non-parametric* survival analysis, eliminating the exponential assumptions, derive bias bounds, and prove asymptotic unbiasedness for the proposed estimators. Moreover, our analysis includes *irregular* sequence lengths, thereby addressing more practical scenarios. We also provide a ready-to-use code of our estimators for practitioners. Supplementary related work is given in App. E.

3. Kaplan-Meier ARL & ADD

We first introduce our notation. For comparison, we then introduce conventional estimators under random length truncation. Finally, we present our proposed estimators.

Definitions. $[n]$ denotes $\{1, 2, \ldots, n\}$ with $n \in \mathbb{N}$. We use P as a probability distribution. $X^{(0,t)} = (X^{(0)}, X^{(1)}, \ldots, X^{(t)})$ denote real-valued random variables representing a sequence of length $t+1$ ($t \in \mathbb{Z}_{\geq 0}$). Each frame $X^{(s)}$ ($s \in \mathbb{Z}_{\geq 0}$) is sampled from the pre-change density $g(X^{(s)} \mid X^{(0,s-1)})$ when $s < \nu$ and from the post-change density $f(X^{(s)} \mid X^{(0,s-1)})$ when $s \geq \nu$, where the changepoint $\nu \in \mathbb{Z}_{\geq 0} \cup \{\infty\}$ is a random variable independent of the observations. $\nu = \infty$ and $\nu = 0$ indicate that all frames in the sequence are sampled from the pre-change and post-change densities, respectively. Let $T \in \mathbb{Z}_{\geq 0}$ denote a random variable representing the sequence length, which is independent of the observations. Let $\tau : X^{(0,T)} \mapsto \tau(X^{(0,T)}) \in \mathbb{Z}_{\geq 0} \cup \{\infty\}$ be a changepoint detector, where $\tau(X^{(0,T)})$ is the detection point of sequence $X^{(0,T)}$. We set $\tau(X^{(0,T)}) = \infty$ when no change is detected within the input sequence of finite length. The ARL and ADD are defined as

$$\mu_\infty := \mathbb{E}[\tau \mid \nu = \infty] \tag{1}$$

$$\text{and } M_\infty := \mathbb{E}[\Delta\tau \mid \Delta\tau \geq 0, \nu < \infty], \tag{2}$$

respectively, where $\Delta\tau := \tau - \nu$. $T = \infty$ is assumed here, and the expectation is also taken over ν. We consider estimation of the ARL and ADD from a dataset $\{(X_i^{(0,T_i)}, \nu_i)\}_{i=1}^N$ with predicted detection points $\{\tau_i\}_{i=1}^N$ from a detector τ, where $N \in \mathbb{N}$ denotes the dataset size, $X_i^{(0,T_i)}$ is a sequence with a random length T_i, ν_i is the corresponding changepoint, and τ_i is the detection point of $X_i^{(0,T_i)}$. Let $\langle \cdot \rangle_{i:\mathcal{C}} := \sum_{i:\mathcal{C}} \cdot / |\{i : \mathcal{C}\}|$ denote the empirical expectation under condition $\mathcal{C}$.

Conventional estimators. A conventional estimator of ARL under random sequence lengths, adapted from (Qiu, 2013) and referred to as the Less-Biased ARL (LB-ARL) in this paper, follows the definition of μ_∞: $\hat{\mu}_T^{(\mathrm{LB})} := \langle \tau_i \rangle_{i:\nu_i=\infty,,\tau_i \leq T_i}$, where only sequences satisfying $\tau_i \leq T_i$ are included. Similarly, a conventional ADD estimator, the LB-ADD, can be defined as $\hat{M}_T^{(\mathrm{LB})} := \langle \Delta\tau_i \rangle_{i:\Delta\tau_i \geq 0, \nu_i < \infty, \Delta\tau_i \leq \Delta T_i}$, where $\Delta T_i := T_i - \nu_i$.

A critical drawback of these conventional estimators is that they ignore sequences in which the QCD model fails to raise an alarm before the horizon ($\tau > T$). Although the LB-ARL has been employed in Monte Carlo simulations of the ARL (Qiu, 2013; Lim & Lee, 2025) when $T = \mathrm{const.} < \infty$, it exhibits a more substantial negative bias than our estimator due to truncation, as proven in Sec.4 and demonstrated in Sec.5.

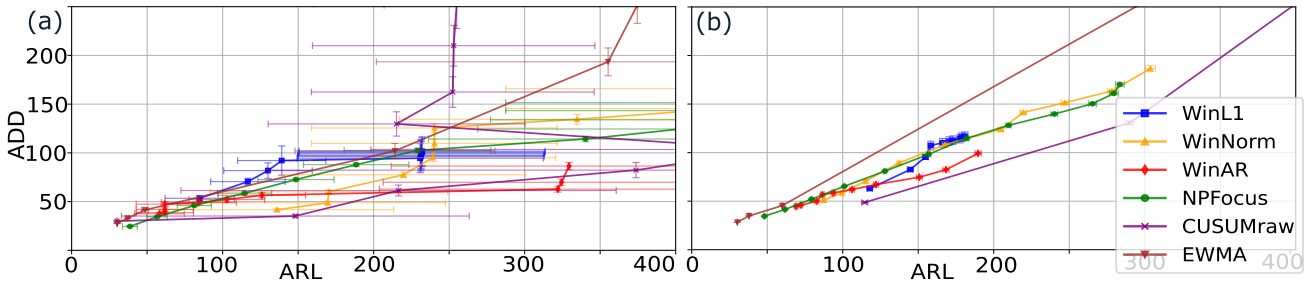


Figure 1. **Evaluation of QCD models on real-world dataset (machine-labeled subset of WISDM Actitracker). (a) LB-ARL & LB-ADD. (b) KM-ARL & KM-ADD.** Error bars represent the standard error of the mean. Our KM-ARL and KM-ADD are more robust to irregular lengths than the conventional estimators (LB-ARL and LB-ADD: see Sec. 3). The figures are shown at the same magnification scale for readability. The complete figure is provided in Fig. 14 in App. C.5. Additional evaluation on a Gaussian process dataset is provided in App. C.3.

Key ideas from survival analysis. To overcome this challenge, we model the detection probability beyond the truncation length, using a non-parametric approach inspired by *survival analysis* (Kleinbaum & Klein, 1996). A central interest in survival analysis is the *survival function* $S(t) := P(\text{event time} > t)$, representing the probability that a patient survives beyond time t ($t \in \mathbb{R}_{\geq 0}$), where the *event time* refers to the time of death. Crucially, the estimation of $S(t)$ on a dataset is performed under *right-censoring*; i.e., the exact event times of several patients are unknown because the patients are lost to follow-up or are still under observation at the end of the study. Thus, we only know the lower bound of the event times of these patients, called *censoring times*. The mean survival time is given by the integral of $S(t)$ over time, i.e., the area under the survival curve, because $\int_0^\infty S(t)dt = \int_0^\infty P(\text{event time} > t)dt = \int_0^\infty \mathbb{E}[\mathbb{1}(\text{event time} > t)]dt = \mathbb{E}[\int_0^\infty \mathbb{1}(\text{event time} > t)dt] = \mathbb{E}[\text{event time}]$, where $\mathbb{1}$ is the indicator function.

KM-ARL. To estimate the ARL under irregular sequence lengths, we draw an analogy by regarding a patient as a sequence, the event time as the detection point, the censoring time as the minimum of the changepoint and the sequence length, and the mean survival time as the ARL. Under this analogy, we estimate the *survival function of detection points* $S^{\mathrm{ARL}}(t) := P(\tau > t \mid \nu = \infty)$ using the Kaplan-Meier estimator (KME) (Kaplan & Meier, 1958), a non-parametric estimator of the survival function (see Eq. (22) in App. A for a simple derivation): $\hat{S}^{\mathrm{ARL}}(t) = \prod_{j:t_j^{\mathrm{ARL}} \leq t}(1 - \frac{d_j^{\mathrm{ARL}}}{n_j^{\mathrm{ARL}}})$, where $0 < t_1^{\mathrm{ARL}} < t_2^{\mathrm{ARL}} < \cdots < t_{N'}^{\mathrm{ARL}}$ are distinct detection points, with $N'_{\mathrm{ARL}} \in \mathbb{N}\,(\leq N)$; $d_j^{\mathrm{ARL}} := |\{i \in [N] \mid \tau_i = t_j^{\mathrm{ARL}}\}|$ is the number of sequences with a detection at t_j^{ARL} ($j \in [N'_{\mathrm{ARL}}]$); and $n_j^{\mathrm{ARL}} := |\{i \in [N] \mid \min\{\tau_i, C_i^{\mathrm{ARL}}\} \geq t_j^{\mathrm{ARL}}\}|$ is the number of sequences neither detected nor censored prior to t_j^{ARL} ($j \in [N'_{\mathrm{ARL}}]$), with the censoring time of sequence i defined as $C_i^{\mathrm{ARL}} := \min\{\nu_i, T_i\}$. An example of $\hat{S}^{\mathrm{ARL}}(t)$ is shown in App. C.1. We propose a non-parametric estimator of the ARL under irregular sequence lengths, termed KM-ARL, as the integral of $\hat{S}^{\mathrm{ARL}}(t)$ over the range $[0, a]$ for arbitrary $a \in \mathbb{R}_{\geq 0}$:

$$\hat{\mu}_T^{(\mathrm{KM})} := \int_0^a \hat{S}^{\mathrm{ARL}}(t)dt. \tag{3}$$

In practice, we set $a = T_{\max} := \max_i\{\min\{\tau_i, C_i^{\mathrm{ARL}}\}\}$, i.e., the maximum last-observed time, following standard practice in survival analysis (Qi & Wang, 2018; Calkins et al., 2018). For a given dataset, choosing $a > T_{\max}$ is irrelevant because it is extrapolation beyond the observed support. Theoretically ideal choices of a are given in Sec. 4. Finally, see App. A for additional motivation and background on our analogy between QCD and survival analysis.

KM-ADD. For the ADD, we regard a patient as a sequence with $\nu_i < \infty$ and $\tau_i \geq \nu_i$, the event time as the detection delay $\Delta\tau_i (\geq 0)$, the censoring time as the sequence length measured from the changepoint ($C_i^{\mathrm{ADD}} := \Delta T_i = T_i - \nu_i$), and the mean survival time as the ADD (see Tab. 1 in App. A for reference). Under this analogy, we propose a non-parametric estimator of the ADD under irregular lengths, termed KM-ADD, as

$$\hat{M}_T^{(\mathrm{KM})} := \int_0^b \hat{S}^{\mathrm{ADD}}(t)dt, \tag{4}$$

where $\hat{S}^{\mathrm{ADD}}(t) := \prod_{j:t_j^{\mathrm{ADD}} \leq t}(1 - \frac{d_j^{\mathrm{ADD}}}{n_j^{\mathrm{ADD}}})$ is a non-parametric estimate of the *survival function of detection delays* $S^{\mathrm{ADD}}(t) := P(\Delta\tau > t \mid \Delta\tau \geq 0, \nu < \infty)$; $0 \leq t_1^{\mathrm{ADD}} < t_2^{\mathrm{ADD}} < \cdots < t_{N'_{\mathrm{ADD}}}^{\mathrm{ADD}}$ are the distinct detection delays, i.e., the sorted unique values of $\Delta\tau_i \geq 0$, with $N'_{\mathrm{ADD}} \in \mathbb{N}\,(\leq N)$; $d_j^{\mathrm{ADD}} := |\{i \in [N] \mid 0 \leq \Delta\tau_i = t_j^{\mathrm{ADD}}\}|$ is the number of sequences with positive detection delay equal to t_j^{ADD} ($j \in [N'_{\mathrm{ADD}}]$); and $n_j^{\mathrm{ADD}} := |\{i \in$

$[N] \mid \min\{\Delta\tau_i, C_i^{\mathrm{ADD}}\} \geq t_j^{\mathrm{ADD}}, \Delta\tau_i \geq 0\}|$ is the number of sequences neither detected nor censored prior to t_j^{ADD} ($j \in [N'_{\mathrm{ADD}}]$). Again, the upper limit b is arbitrary, and we set $b = \Delta T_{\max} := \max_i\{\min\{\Delta\tau_i, C_i^{\mathrm{ADD}}\} \mid \Delta\tau_i \geq 0, \nu_i < \infty\}$ in our experiments. Finally, we provide a summary of our analogy between QCD and survival analysis in Tab. 1 in App. A.

4. Bias Analysis

We derive bias bounds for the KM-ARL and KM-ADD and prove that they are asymptotically unbiased unless extrapolation is required. We first examine the estimation bias of the KM-ARL:

$$\mathcal{B}(\hat{\mu}_T^{(\mathrm{KM})}) := \mathbb{E}[\hat{\mu}_T^{(\mathrm{KM})}] - \mu_\infty, \tag{5}$$

where $\mu_\infty := \mathbb{E}[\tau \mid \nu = \infty] < \infty$ is the true ARL under infinite sequence length (we assume all relevant finiteness hereafter). We decompose this bias into two components: the *finite-sample bias* and the *truncation bias*

$$\mathcal{B}_{\mathrm{FS}}(\hat{\mu}_T^{(\mathrm{KM})}) := \mathbb{E}[\hat{\mu}_T^{(\mathrm{KM})}] - \mu_T^{(\mathrm{KM})} \tag{6}$$

$$\mathcal{B}_{\mathrm{TR}}(\hat{\mu}_T^{(\mathrm{KM})}) := \mu_T^{(\mathrm{KM})} - \mu_\infty^{(\mathrm{KM})} \tag{7}$$

so that $\mathcal{B}(\hat{\mu}_T^{(\mathrm{KM})}) = \mathcal{B}_{\mathrm{FS}}(\hat{\mu}_T^{(\mathrm{KM})}) + \mathcal{B}_{\mathrm{TR}}(\hat{\mu}_T^{(\mathrm{KM})})$, where $\mu_T^{(\mathrm{KM})} := \int_0^a S_{\mathrm{ARL}}(t)dt$ is the KM-ARL with the true survival function of detection points.

Finite-sample bias. $\mathcal{B}_{\mathrm{FS}}(\hat{\mu}_T^{(\mathrm{KM})})$ quantifies the error arising from a finite dataset of sequences with finite and irregular lengths and dominates the total bias when the dataset size is small. We derive a bound for $\mathcal{B}_{\mathrm{FS}}(\hat{\mu}_T^{(\mathrm{KM})})$, indicating that it decays exponentially as $N \to \infty$:

Theorem 4.1 (Finite-sample bias bounds for KM-ARL)**.** *We idealize the time index as continuous without loss of generality, for technical convenience. Let $F^{\mathrm{ARL}}, G^{\mathrm{ARL}}$, and H^{ARL} be the cumulative distribution functions (CDFs) of τ, C^{ARL}, and $\min\{\tau, C^{\mathrm{ARL}}\}$, respectively. Assume that (i) F^{ARL} and G^{ARL} do not have common discontinuities, (ii) τ and C^{ARL} are independent, known as the independent censoring (or non-informative censoring) assumption (Ranganathan & Pramesh, 2012), and (iii) H^{ARL} is continuous. Then, we have for any $a \in \mathbb{R}_{\geq 0}$*

$$-\int_0^a t\, G^{\mathrm{ARL}}(t)\, H^{\mathrm{ARL}}(t)^{N-1}\, dF^{\mathrm{ARL}}(t) \leq \mathcal{B}_{\mathrm{FS}}(\hat{\mu}_T^{(\mathrm{KM})}) \leq \int_0^a a\, G^{\mathrm{ARL}}(t)\, H^{\mathrm{ARL}}(t)^{N-1}\, dF^{\mathrm{ARL}}(t). \tag{8}$$

Our proof is based on (Stute & Wang, 1993; Stute, 1994) and given in App. B.1. Assumption (i) is a technical requirement and is valid unless pathological situations are considered. Assumption (ii) holds for online QCD models that do not look ahead the input sequence, which are the focus in this paper. This is because: detection points τ_i prior to censoring can be regarded as samples from $P(\tau \mid \nu = \infty)$; $P(\tau \mid \nu = \infty)$, $P(\nu)$, and $P(T)$ are independent; and thus, $P(\tau \mid \nu = \infty)$ and $P(C := \min\{\nu, T\})$ are also independent. In contrast, Assumption (ii) does not hold for offline changepoint detection models because τ depends on $X^{(0,T)}$, which in turn depends on ν and T. Assumption (iii) is also a technical requirement.

According to Thm. 4.1, the finite-sample bias exhibits the following properties. First, it decays exponentially to zero as $N \to \infty$ if $H^{\mathrm{ARL}}(t) < 1$ for $t \in [0, a]$. Second, it vanishes when $G^{\mathrm{ARL}} = 0$, i.e., in the absence of censoring (no changepoints or horizons) in $t \in [0, a]$, which is desirable because estimation becomes more accurate as censoring decreases. Third, the bound becomes looser for larger a because G^{ARL} and H^{ARL} approach 1 monotonically as t increases. However, the truncation bias decreases for larger a, leading to a tradeoff between the finite-sample and truncation biases. Finally, we note that empirically verifying the convergence of the finite-sample bias is challenging because, for small sample sizes, estimation variance overshadows the bias.

We can derive a similar bound for the finite-sample bias of the KM-ADD (proof is given in App. B.2), ensuring that it also decays exponentially as $N'' \to \infty$, where N'' is the number of sequences with $\nu_i < \infty$ and $\Delta\tau_i \geq 0$:

Theorem 4.2 (Finite-sample bias bounds for KM-ADD)**.** *We idealize the time index as continuous without loss of generality, for technical convenience. Consider only sequences with $\nu_i < \infty$ and $\Delta\tau_i \geq 0$ in the dataset. Let $F^{\mathrm{ADD}}, G^{\mathrm{ADD}}$, and H^{ADD} be the CDFs of $\Delta\tau$, C^{ADD}, and $\min\{\Delta\tau, C^{\mathrm{ADD}}\}$, respectively. Assume that (i) F^{ADD} and G^{ADD} do not have common discontinuities, (ii) $\Delta\tau$ and C^{ADD} are independent, and (iii) H^{ADD} is continuous. Then, for any $b \in \mathbb{R}_{\geq 0}$,*

$$-\int_0^b t\, G^{\mathrm{ADD}}(t)\, H^{\mathrm{ADD}}(t)^{N''-1}\, dF^{\mathrm{ADD}}(t) \leq \mathcal{B}_{\mathrm{FS}}(\hat{M}_T^{(\mathrm{KM})}) \leq \int_0^b b\, G^{\mathrm{ADD}}(t)\, H^{\mathrm{ADD}}(t)^{N''-1}\, dF^{\mathrm{ADD}}(t). \tag{9}$$

Here, we defined the finite-sample bias of the KM-ADD similarly to that of the KM-ARL: $\mathcal{B}_{\mathrm{FS}}(\hat{M}_T^{(\mathrm{KM})}) := \mathbb{E}[\hat{M}_T^{(\mathrm{KM})}] - M_T^{(\mathrm{KM})}$, where $M_T^{(\mathrm{KM})} := \int_0^b S^{\mathrm{ADD}}(t)dt$ is the KM-ADD with the true survival function of detection delays. Assumption (i) and (iii) are technical conditions, as previously noted below Thm. 4.1. Assumption (ii) is the independent censoring assumption for the KM-ADD, which can be justified in online QCD models by the approx-

imation $P(\Delta\tau \mid \Delta\tau \geq 0, \nu < \infty) \approx P(\tau \mid \nu = 0)$ and $P(C := T - \nu \mid \Delta\tau \geq 0, \nu < \infty) \approx P(C := T \mid \nu = 0)$ because $P(\tau \mid \nu = 0)$ and $P(C := T \mid \nu = 0)$ are independent. This approximation implies that the distributions of the detection delay $\tau - \nu$ and the censoring time $T - \nu$ are approximately equal to their respective distributions measured from $t = 0$ rather than from ν. In App. F.1, we further quantify violations of the independent censoring assumption and provide its potential remedies.

Truncation bias. $\mathcal{B}_{\mathrm{TR}}(\hat{\mu}_T^{(\mathrm{KM})})$ captures the error from sequence truncation, which dominates the total bias when sequence lengths are short. It vanishes as a increases because $\mathcal{B}_{\mathrm{TR}}(\hat{\mu}_T^{(\mathrm{KM})}) = \int_0^a S^{\mathrm{ARL}}(t)dt - \mathbb{E}[\tau \mid \nu < \infty] = -\int_a^\infty S^{\mathrm{ARL}}(t) \to 0$, where we used $\mathbb{E}[\tau \mid \nu < \infty] = \int_0^\infty S^{\mathrm{ARL}}(t)dt$. The convergence rate depends on the underlying distributions. In Thm. 4.3 below, we derive a bound for $\mathcal{B}_{\mathrm{TR}}(\hat{\mu}_T^{(\mathrm{KM})})$, indicating that it is non-positive and minor than that of the conventional LB-ARL. In Sec. 5, we will empirically justify our result (Fig. 2). Proof is given in App. B.3. Note that given an evaluation dataset, we can easily specify $T^*_{\max}$ defined below (App. F).

Theorem 4.3 (Truncation bias bound for KM-ARL)**.** *Define the truncation bias of the LB-ARL as* $\mathcal{B}_{\mathrm{TR}}(\hat{\mu}_T^{(\mathrm{LB})}) := \mu_T^{(\mathrm{LB})} - \mu_\infty$, *where* $\mu_T^{(\mathrm{LB})} := \mathbb{E}[\tau \mid \nu = \infty, \tau \leq T]$ *is the true ARL under random lengths. For* $a = T^*_{\max}$, *where* $T^*_{\max} := \inf\{T \mid \mathrm{CDF}(T) = 1\}$ *is the least upper bound for the support of the CDF of* T, *we have* $\mathcal{B}_{\mathrm{TR}}(\hat{\mu}_T^{(\mathrm{LB})}) \leq \mathcal{B}_{\mathrm{TR}}(\hat{\mu}_T^{(\mathrm{KM})}) \leq 0$.

A similar bound holds for the KM-ADD. Proof is given in App. B.4.

Theorem 4.4 (Truncation bias bound for KM-ADD)**.** *Define the truncation bias of the LB-ADD as* $\mathcal{B}_{\mathrm{TR}}(\hat{M}_T^{(\mathrm{LB})}) := M_T^{(\mathrm{LB})} - M_\infty$, *where* $M_T^{(\mathrm{LB})} := \mathbb{E}[\Delta\tau \mid \nu < \infty, 0 \leq \Delta\tau \leq \Delta T]$ *is the true ADD under random sequence lengths. For* $b = \Delta T^*_{\max}$, *where* $\Delta T^*_{\max} := \inf\{\Delta T := T - \nu \mid \mathrm{CDF}(\Delta T) = 1\}$, *we have* $\mathcal{B}_{\mathrm{TR}}(\hat{M}_T^{(\mathrm{LB})}) \leq \mathcal{B}_{\mathrm{TR}}(\hat{M}_T^{(\mathrm{KM})}) \leq 0$, *under the independent censoring assumption in Thm. 4.2.*

Total Estimation Bias. Finally, we examine the total estimation bias of the KM-ARL: $\mathcal{B}(\hat{\mu}_T^{(\mathrm{KM})}) = \mathcal{B}_{\mathrm{FS}}(\hat{\mu}_T^{(\mathrm{KM})}) + \mathcal{B}_{\mathrm{TR}}(\hat{\mu}_T^{(\mathrm{KM})})$ (a similar discussion below holds for the KM-ADD). From Thm. 4.1 and $\mathcal{B}_{\mathrm{TR}}(\hat{\mu}_T^{(\mathrm{KM})}) = -\int_a^\infty S^{\mathrm{ARL}}(t)$, we have for any $a \in \mathbb{R}_{\geq 0}$

$$-\int_0^a t\, G^{\mathrm{ARL}}(t)\, H^{\mathrm{ARL}}(t)^{N-1}\, dF^{\mathrm{ARL}}(t) - \int_a^\infty S^{\mathrm{ARL}}(t) \leq \mathcal{B}(\hat{\mu}_T^{(\mathrm{KM})}) \leq \int_0^a a\, G^{\mathrm{ARL}}(t)\, H^{\mathrm{ARL}}(t)^{N-1}\, dF^{\mathrm{ARL}}(t) - \int_a^\infty S^{\mathrm{ARL}}(t). \tag{10}$$

Define $t_F := \inf\{t \in \mathbb{R}_{\geq 0} \mid F^{\mathrm{ARL}}(t) = 1\}$ and $t_H := \inf\{t \in \mathbb{R}_{\geq 0} \mid H^{\mathrm{ARL}}(t) = 1\}$, and set a to t_F. Then, since $S^{\mathrm{ARL}} = 1 - F^{\mathrm{ARL}}$, we have $\int_a^\infty S^{\mathrm{ARL}}(t) = 0$; i.e., the truncation bias vanishes. Therefore,

$$-\int_0^{t_F} t\, G^{\mathrm{ARL}}(t)\, H^{\mathrm{ARL}}(t)^{N-1}\, dF^{\mathrm{ARL}}(t) \leq \mathcal{B}(\hat{\mu}_T^{(\mathrm{KM})}) \leq \int_0^{t_F} t_F\, G^{\mathrm{ARL}}(t)\, H^{\mathrm{ARL}}(t)^{N-1}\, dF^{\mathrm{ARL}}(t). \tag{11}$$

If $t_F < t_H$, the bound decays exponentially as $N \to \infty$ because $H^{\mathrm{ARL}}(t) < 1$ for $t \in [0, t_F]$; i.e., *the KM-ARL is asymptotically unbiased if* $t_F < t_H$. Otherwise, there are non-vanishing terms $-\int_{t_H}^{t_F} t\, G^{\mathrm{ARL}}(t)\, dF^{\mathrm{ARL}}(t)$ and $\int_{t_H}^{t_F} t_F\, G^{\mathrm{ARL}}(t)\, dF^{\mathrm{ARL}}(t)$ in the lower and upper bounds, respectively. This is reasonable because $t_F \geq t_H$ implies that detection points may occur beyond censoring times with non-zero probability; i.e., not all detection points are observable. Consequently, estimating the ARL without bias or additional assumptions is impossible, necessitating extrapolation.

5. Experiment

To demonstrate the practical relevance of our KM-ARL and KM-ADD for sequences with limited and irregular lengths, we use both simulated and real-world datasets. Our results show that these metrics reduce estimation bias compared to baseline estimators, enhance robustness to such sequences, and thereby improve interpretability and facilitate empirical, intuitive model selection.

Datasets. We use two simulation datasets, the Gaussian and Poisson processes, and one real-world dataset, the WISDM Actitracker (Kwapisz et al., 2011). The pre-change and post-change Gaussian processes have the mean of 0 and 0.1, respectively, with preserving the variance equal to 0.1. We use two types of changepoint distributions: geometric and uniform. Several sequences have no changepoint, and the number of with-change sequences depends on the changepoint distributions. We also simulate irregular length by randomly truncating the sequences. The setup and experimental results for Poisson processes are provided in App. C.4 due to page limitations. The results are similar

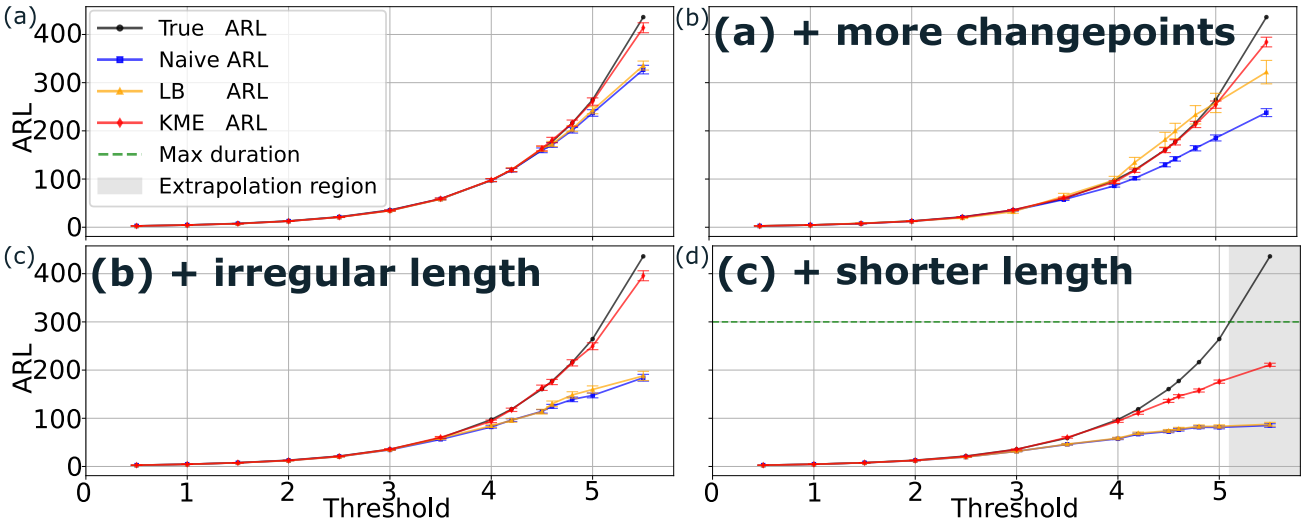


Figure 2. **Threshold of detector vs. ARL.** The KM-ARL provides more accurate estimates of the true ARL even when sequences contain changepoints and have limited and irregular lengths. The Gaussian process dataset contains 1000 sequences. The GSR with ground-truth statistics is evaluated under various thresholds. Changepoints are sampled uniformly. Error bars represent the standard error of the mean. **(a) Sequence length is** 1000**, with** 10% **of sequences containing a changepoint. (b) Sequence length is** 1000**, with** 90% **of sequences containing a changepoint. (c) Sequence lengths vary irregularly in the range** $[100, 1000]$**, with** 90% **of sequences containing a changepoint. (d) Sequence lengths vary irregularly in the range** $[30, 300]$**, with** 90% **of sequences containing a changepoint.** In (d), $\mathrm{ARL} > 300$ (gray area) is an extrapolation region (excluding the true ARL).

to those obtained for the Gaussian processes. The WISDM Actitracker is a large, real-world dataset for smartphone-based human activity recognition (walking, jogging, stairs, sitting, standing, and lying down) collected by the WISDM Lab at Fordham University, offering both user-labeled and machine-labeled data. The labels specify activities and their temporal intervals. We provide the results for the machine-labeled subset in the main text, and the results for the user-labeled subset are provided in Fig. 15 in App. C.5. The machine-labeled subset contains 51,326 sequences, and the sequence lengths exhibit substantial irregularity from 1 to 54,401 after our preprocesses. See App. D.3 for more statistical information of the WISDM Actitracker dataset. We remove the sequences with length 1, as they are not informative when computing performance metrics of QCD models. The preprocesses are detailed in our code and App. D.2. Additionally, relevance of our setting, namely, the requirement for datasets containing multiple sequences with changepoint labels is further clarified in App. F.2 for interested readers.

5.1. Estimation on Simulation Dataset

Detection models. The QCD model used for the simulation datasets is the generalized Shiryaev-Roberts (GSR) procedure with ground-truth statistics. The GSR raises an alarm when the following statistic hits a pre-defined threshold: $R(t) = \omega \prod_{s=0}^{t} \mathcal{L}(s) + \sum_{k=0}^{t} \prod_{s=k}^{t} \mathcal{L}(s)$, where the likelihood ratio is denoted by $\mathcal{L}(t) = f\left(X^{(t)} \mid X^{(0,t-1)}\right) / g\left(X^{(t)} \mid X^{(0,t-1)}\right)$, and ω controls the strength of warm start and is set to 0 in our experiments. Additionally, we provide the results for the cumulative sum (CUSUM) procedure in App. C.6.

True ARLs & ADDs. For the experiments on the Gaussian and Poisson process datasets, we simulate the ground-truth ARLs and ADDs by generating sequences of effectively infinite length. Their error bars are omitted because all errors are sufficiently small, with relative errors $\lesssim 10^{-3}$.

Naive ARL. We introduce another baseline metric, referred to as the Naive ARL. While the LB-ARL use only sequences with $\nu_i = \infty$ and $\tau_i \leq T_i$ for computing the ARL, the Naive ARL utilizes sequences with $\tau_i < \nu_i$ and $\tau_i \leq T_i$: $\hat{\mu}_T^{(\mathrm{NV})} := \langle \tau_i \rangle_{i:\tau_i<\nu_i,\tau_i\leq T_i}$. The number of sequences used for the Naive ARL is larger than that of the LB-ARL because the condition $\tau_i < \nu_i \wedge \tau_i \leq T_i$ includes $\nu_i = \infty \wedge \tau_i \leq T_i$; however, the Naive ARL has a non-vanishing bias because $\mathbb{E}[\hat{\mu}_T^{(\mathrm{NV})}] = \mathbb{E}[\tau \mid \tau < \nu, \tau < T]$ under minor assumptions, which is not equal to $\mathbb{E}[\tau \mid \nu = \infty, \tau < T]$ or $\mathbb{E}[\tau \mid \nu = \infty]$.

Result on ARL. Fig. 2 presents the true ARL, Naive ARL, LB-ARL, and KM-ARL across various GSR thresholds, demonstrating that the KM-ARL provides more accurate estimates of the true ARL even when sequences contain changepoints and have limited and irregular lengths. Fig. 2(a) simulates an ideal light-censoring scenario, where

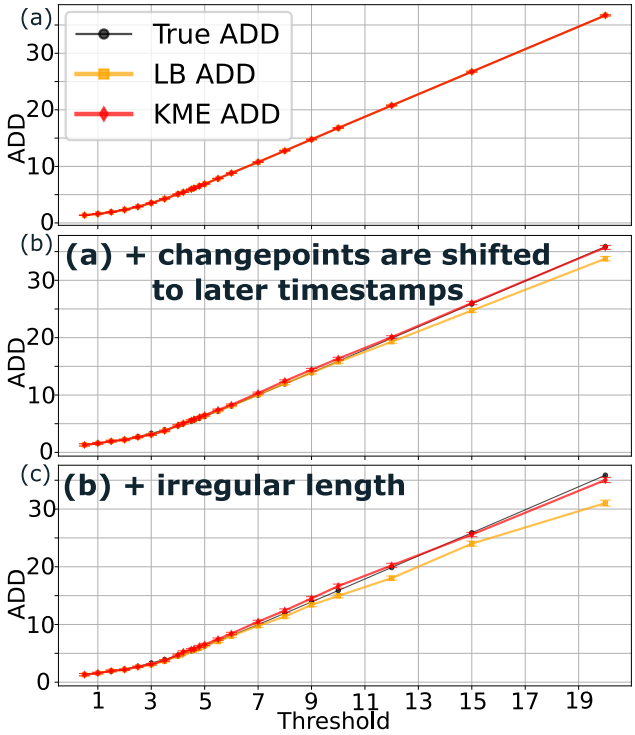


Figure 3. **Threshold of detector vs. ADD.** The KM-ADD provides more accurate estimates of the true ADD even when sequence lengths are limited and irregular. The Gaussian process dataset has 10000 sequences. Changepoints are sampled from the geometric distribution with the success probability $p \in [0, 1]$. A smaller p leads to more sparse and delayed changepoints. Other settings are identical to that of Fig. 2. **(a) Sequence length is** 100**, with** $p = 0.25$**. (b) Sequence length is** 100**, with** $p = 0.001$**.** The changepoints are shifted to later timestamps, decreasing the chance of detection due to the length limit. **(c) Sequence lengths vary irregularly in the range** $[10, 100]$**, with** $p = 0.001$**.**

most detection points occur before changepoints or truncation. The true ARL is less than $T_{\max}$, the sequence length T is constant (=1000), and only 10% of the sequences contain a changepoint. Fig. 2(b) simulates a heavier-censoring scenario, where 90% of the sequences contain a changepoint. Fig. 2(c) simulates an even more heavily censored scenario, where, in addition to Fig. 2(b), sequence lengths are irregular and randomly sampled from $[100, 1000]$. Our KM-ARL remains robust under this condition, utilizing with-change and truncated sequences, unlike the Naive ARL and the LB-ARL. Fig. 2(d) simulates a challenging and realistic scenario, where, in addition to Fig. 2(c), $T_{\max}$ is reduced to 300, and sequence lengths are sampled from $[30, 300]$. The gray area indicates $\mathrm{ARL} > T_{\max} = 300$, where no data are observed, i.e., an extrapolation region (excluding the true ARL). Bias becomes non-negligible in $\mathrm{ARL} \gtrsim T_{\max}$ due to truncation. In this region, no metric can reliably estimate the ARL without bias or additional assumptions, as noted Sec. 4. To further mitigate truncation bias, one may combine our estimators with parametric extrapolation methods for the survival function, such as those proposed in (Sahki et al., 2020).

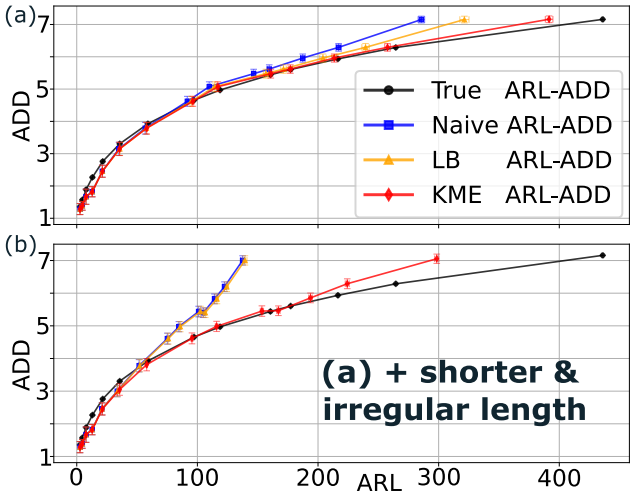


Figure 4. **ARL-ADD tradeoff curves.** The KM-ARL and KM-ADD provide more accurate estimates of the true ARL-ADD curve even when sequences contain changepoints and have limited and irregular lengths. The setup is identical to that of Fig. 2, except that the dataset size is 10000. 50% of sequences contain a changepoint. The LB-ADD is used for the Naive ARL's ADD. **(a) Sequence length is** 1000**. (b) Sequence lengths vary irregularly in the range** $[50, 500]$**.** The complete ablation study is given in App. C.2.

Result on ADD. Fig. 3 presents the true ADD, LB-ADD, and KM-ADD across various GSR thresholds, demonstrating that the KM-ADD provides more accurate estimates of the true ADD even when sequences have limited and irregular lengths. Fig. 3(a) simulates an ideal light-censoring scenario, where most detection points occur before truncation. T is constant and set to 100, and $\Delta T_{\max}$ is also set to 100, which is greater than the true ADD. Fig. 3(b) simulates a heavier-censoring scenario, where changepoints are shifted to later timestamps, reducing the chance of detection before truncation. Fig. 3(c) simulates a challenging and realistic scenario, where, in addition to Fig. 3(b), sequence length are irregular and sampled from $[10, 100]$, further reducing the chance of detection before truncation. The KM-ADD still remains robust, utilizing truncated sequences, unlike the conventional LB-ADD.

ARL-ADD tradeoff curve. Fig. 4 shows ARL-ADD tradeoff curves, which are theoretically related to the optimality of QCD models and are used for model evaluation in practice. Fig. 4 demonstrates that the KM-ARL and KM-ADD can more accurately estimate the ARL-ADD curve

than the others. Fig. 4(a) simulates a light-censored scenario, where most detection points occur before censoring. The sequence length T is constant and set to $T = T_{\max} = 500$. Fig. 4(b) simulates a challenging and realistic scenario, where censoring is heavier. Sequence lengths are irregular and sampled from $[50, 500]$. The KM-ARL and KM-ADD remains more robust than the baselines. Again, bias is non-negligible in the region $\mathrm{ARL} \approx T_{\max} = 500$, where extrapolation region is nearby.

5.2. Evaluation of Models on Real-world Dataset

We further evaluate six QCD models on a real-world, challenging dataset: the WISDM Actitracker (Kwapisz et al., 2011), which contains 51,326 sequences with substantial irregularity of lengths from 1 to 54,401. The computational costs for KM-ARL and KM-ADD are negligible in our experiments compared with the time required to run the QCD algorithms.

QCD models. We use the following online QCD models, including window-based, frame-based, parametric, and non-parametric models: Window L1 (Bai, 1995), Window Normal (Lavielle, 1999; Lavielle & Teyssiere, 2006), Window AR (Bai, 2000), non-parametric focused changepoint detection (NP-FOCuS) (Romano et al., 2024a), CUSUM (Page, 1954), and exponentially weighted moving average (EWMA) (Roberts, 1959). The window size and burn-in interval are fixed at 30, ensuring much smaller than the average length. Most other hyperparameters are left at default values. See App. D.1 for details of the models and their hyperparameters.

Result. The ARL-ADD tradeoff curves are presented in Fig. 1, demonstrating that our KM-ARL and KM-ADD are more robust even when censoring is significantly heavy. Fig. 1(a) shows that LB-ARL and LB-ADD become unstable and exhibit high variance, particularly at large QCD thresholds. This instability arises because the detector often fails to raise an alarm within the sequence length, and the limited number of sequences further amplifies the empirical variance. In contrast, Fig. 1(b) shows that our KME-based estimators do not suffer from this issue. These metrics reduce variance compared to baseline estimators because they are calculated from a constant number of sequences, regardless of the threshold, which enhances their robustness against censoring. Therefore, the KM-ARL and KM-ADD improve interpretability and facilitating empirical, intuitive model selection. See Apps. D.4 & F for more about empirical computation and theoretical estimation of variance, respectively.

6. Discussion, Limitations, and Future Work

To further reduce estimation bias. For further bias reduction, one may use the bootstrap method (Efron & Tibshirani, 1994) to mitigate the finite-sample bias, although this effect is typically overshadowed by the truncation bias. To address the truncation bias beyond extrapolation, a parametric estimator can be combined with our estimators, as in (Sahki et al., 2020); however, note that the estimation accuracy deteriorates when the parametric assumption is invalid.

Changepoint isolation and multiple changepoint detection. Our work assumes a single changepoint type; however, changepoint isolation (classification) (Tartakovsky et al., 2014) is often required in practice. We conjecture that our survival analysis–based approach can be extended to this challenging setting. A promising direction is to use survival models under *competitive risks* (Morita, 2021), treating different changepoint types as distinct causes of death. Similarly, while our method does not currently support multiple changepoint detection (Niu et al., 2016), it may leverage *multi-state models under competing risks* (Therneau et al., 2020; Beyersmann et al., 2011; Mills, 2011), which are well established in survival analysis.

Probability of false alarms. The probability of false alarms (PFA) is another standard metric in QCD theory, commonly used to establish Bayesian optimality (Tartakovsky, 2019). However, it is also affected by right-censoring in real-world scenarios. We hypothesized that multi-state models under competitive risks, such as the Aalen–Johansen estimator (AJE) (Aalen & Johansen, 1978), could alleviate this issue. Our preliminary experiments, however, failed to accurately estimate PFAs, possibly because the AJE assumes no event ties, an assumption often violated when continuous-time sequences are discretized into frames. This finding suggests that more careful approaches, such as those from discrete-time survival analysis (Tutz et al., 2016), are required.

Dependent censoring. In Sec. 4, we assume independent censoring for KM-ADD, supported by the distributional approximation therein. This assumption can potentially be relaxed by leveraging extensive research on dependent censoring in survival analysis (Hsu & Taylor, 2010; Lin et al., 2023; Crommen et al., 2025). Doing so would not only eliminate the assumption but also extend our bias analysis to offline QCD, where censoring depends on detection. See App. F.1 for more details.

Heavy censoring. We do not recommend using KM-ARL or KM-ADD when datasets are severely imbalanced between pre-change and post-change sequences because this leads to significantly heavy censoring and inflated

finite-sample and truncation bias. While the extent of acceptable censoring depends on the dataset and underlying distributions, Malmquist (2025) reports that, for 30, 50, 100, or 150 subjects (sequences), censoring rates up to 90% are tolerable if censoring is uniform. However, for dependent censoring occurring just before event time (detection), estimation bias increases sharply (see App. F.1.1). Nonetheless, several studies have proposed modifications to the KME to address heavy censoring (Shafiq et al., 2007; Zare & Mahmoodi, 2013), which can be integrated with our KM-ARL and KM-ADD.

Acknowledgements

We thank Akira Kitaoka for his insightful and detailed comments on our work. We also thank all reviewers of our manuscript for their feedback.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Aalen, O. O. and Johansen, S. An empirical transition matrix for non-homogeneous Markov chains based on censored observations. *Scandinavian journal of statistics*, pp. 141–150, 1978.

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. License: Apache License 2.0. Software available from tensorflow.org.

Ahmed, N., Hasan, M. M., Babu, M. A., Ahmmed, M. M., Rahman, M. M., and Mahmud, M. DeepLLR-CUSUM: Sequential change detection with learned log-likelihood ratios for site reliability engineering. In *5th Muslims in ML Workshop co-located with NeurIPS 2025*, 2025. URL https://openreview.net/forum?id=XXpt1NHd4B.

Akritas, M. G. The central limit theorem under censoring. *Bernoulli*, 6(6):1109–1120, 2000. ISSN 13507265. URL http://www.jstor.org/stable/3318473.

Alakent, B. and Mutlu, E. C. Application of robust estimators in Shewhart S-charts. *arXiv preprint arXiv:1812.11132*, 2018.

Arabie, P. and Boorman, S. A. Multidimensional scaling of measures of distance between partitions. *Journal of Mathematical Psychology*, 10(2):148–203, 1973.

Arbelaez, P., Maire, M., Fowlkes, C., and Malik, J. Contour detection and hierarchical image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 33(5):898–916, 2010.

Areepong, Y. and Peerajit, W. Integral equation solutions for the average run length for monitoring shifts in the mean of a generalized seasonal ARFIMAX$(p, d, q, r)_s$ process running on a CUSUM control chart. *Plos one*, 17 (2):e0264283, 2022.

Bai, J. Least absolute deviation estimation of a shift. *Econometric Theory*, 11(3):403–436, 1995.

Bai, J. Vector autoregressive models with structural changes in regression coefficients and in variance-covariance matrices. Technical report, China Economics and Management Academy, Central University of Finance and Economics, 2000.

Beyersmann, J., Allignol, A., and Schumacher, M. *Competing risks and multistate models with R*. Springer Science & Business Media, 2011.

Bishop, C. M. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006. ISBN 0387310738.

Biswas, P. and Kalbfleisch, J. D. A risk-adjusted cusum in continuous time based on the cox model. *Statistics in medicine*, 27(17):3382–3406, 2008.

Borror, C., {Bert Keats}, J., and Montgomery, D. Robustness of the time between events CUSUM. *International Journal of Production Research*, 41(15):3435–3444, October 2003. ISSN 0020-7543. doi: 10.1080/0020754031000138321.

Bouchikhi, I., Ferrari, A., Richard, C., Bourrier, A., and Bernot, M. Kernel based online change point detection. In *2019 27th European Signal Processing Conference (EUSIPCO)*, pp. 1–5. IEEE, 2019.

Bradley, T., Alhajjar, E., and Bastian, N. D. Novelty detection in network traffic: Using survival analysis for feature identification. In *2023 IEEE International Conference on Assured Autonomy (ICAA)*, pp. 11–18, 2023. doi: 10.1109/ICAA58325.2023.00010.

Calkins, K. L., Canan, C. E., Moore, R. D., Lesko, C. R., and Lau, B. An application of restricted mean survival time in a competing risks setting: comparing time to art initiation by injection drug use. *BMC medical research methodology*, 18(1):27, 2018.

Cox, D. R. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972. ISSN 00359246. URL http://www.jstor.org/stable/2985181.

Crommen, G., Deresa, N. W., D'Haen, M., Ding, J., Willems, I., and Van Keilegom, I. Recent advances in copula-based methods for dependent censoring. *SORT-Statistics and Operations Research Transactions*, pp. 3–42, 2025.

Crowder, S. v. A simple method for studying run–length distributions of exponentially weighted moving average charts. *Technometrics*, 29(4):401–407, 1987.

Davidson-Pilon, C. lifelines: survival analysis in Python. *Journal of Open Source Software*, 4(40):1317, 2019.

Eagle, N. and Pentland, A. Reality mining: sensing complex social systems. *Personal and ubiquitous computing*, 10 (4):255–268, 2006.

Ebrahimi, N. and Molefe, D. Survival function estimation when lifetime and censoring time are dependent. *Journal of multivariate analysis*, 87(1):101–132, 2003.

Efron, B. and Tibshirani, R. J. *An introduction to the bootstrap*. Chapman and Hall/CRC, 1994.

Everingham, M., Van Gool, L., Williams, C. K., Winn, J., and Zisserman, A. The PASCAL visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

Fu, J. C., Spiring, F. A., and Xie, H. On the average run lengths of quality control schemes using a Markov chain approach. *Statistics & Probability Letters*, 56(4):369–380, 2002.

GANDY, A., KVALØY, J. T., BOTTLE, A., and ZHOU, F. Risk-adjusted monitoring of time to event. *Biometrika*, 97(2):375–388, 2010. ISSN 00063444, 14643510. URL http://www.jstor.org/stable/25734092.

Gierz, K. *Inference and Estimation in Change Point Models for Censored Data*. PhD thesis, Old Dominion University, 2020.

Gomon, D., Putter, H., Nelissen, R. G., and Van Der Pas, S. CGR-CUSUM: a continuous time generalized rapid response cumulative sum chart. *Biostatistics*, 25(1):253–269, 2024.

Gopalan, A., Lakshminarayanan, B., and Saligrama, V. Bandit quickest changepoint detection. *Advances in Neural Information Processing Systems*, 34:29064–29073, 2021.

Hadjiliadis, O., Schaefer, T., and Poor, H. V. Quickest detection in coupled systems. In *Proceedings of the 48th IEEE Conference on Decision and Control (CDC) held jointly with 2009 28th Chinese Control Conference*, pp. 4723–4728. IEEE, 2009.

Haq, A. and Woodall, W. H. A note on an average run length calculation for the EWMA and other charts. *Quality and Reliability Engineering International*, 38(8):4351–4355, 2022.

Haris, B. C., Pradhan, G., Misra, A., Prasanna, S., Das, R. K., and Sinha, R. Multivariability speaker recognition database in indian scenario. *International Journal of Speech Technology*, 15(4):441–453, 2012.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., Del Río, J. F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., and Oliphant, T. E. Array programming with NumPy. *Nature*, 585(7825):357–362, 09 2020. License: BSD 3-Clause "New" or "Revised" License.

Hawkins, D. M., Qiu, P., and Kang, C. W. The changepoint model for statistical process control. *Journal of quality technology*, 35(4):355–366, 2003.

Heilbron, F. C., Escorcia, V., Ghanem, B., and Niebles, J. C. Activitynet: A large-scale video benchmark for human activity understanding. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 961–970, June 2015. doi: 10.1109/CVPR.2015.7298698.

Hernàndez-Orallo, J., Ferri, C., Lachiche, N., and Flach, P. The 1st workshop on ROC analysis in artificial intelligence (ROCAI-2004). *ACM SIGKDD Explorations Newsletter*, 6(2):159–161, 2004.

Hsu, C.-H. and Taylor, J. M. A robust weighted Kaplan-Meier approach for data with dependent censoring using linear combinations of prognostic covariates. *Statistics in medicine*, 29(21):2215–2223, 2010.

Huang, Y.-H. and Veeravalli, V. V. High probability latency quickest change detection over a finite horizon. In *2024 IEEE International Symposium on Information Theory (ISIT)*, pp. 1047–1052. IEEE, 2024.

Hubert, L. and Arabie, P. Comparing partitions. *Journal of classification*, 2(1):193–218, 1985.

Jiang, Y.-G., Liu, J., Roshan Zamir, A., Toderici, G., Laptev, I., Shah, M., and Sukthankar, R. THUMOS challenge: Action recognition with a large number of classes. http://crcv.ucf.edu/THUMOS14/, 2014.

Johnson, P. and Pedersen, J. L. Bayesian changepoint detection for epidemic models. *Scientific Reports*, 15(1): 20545, 2025.

Kaplan, E. L. and Meier, P. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481, 1958.

Kei, Y. L., Li, J., Li, H., Chen, Y., and PADILLA, O. H. M. Change point detection in dynamic graphs with decoder-only latent space model. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=DVeFqV56Iz.

Khamesi, V. ocpdet: A Python package for online changepoint detection in univariate and multivariate data, October 2022. URL https://doi.org/10.5281/zenodo.7632721.

Klein, J. and Moeschberger, M. Asymptotic bias of the product limit estimator under dependent competing risks. *Indian Journal of Productivity, Reliability and Quality Control*, 9:1–7, 1984.

Kleinbaum, D. G. and Klein, M. *Survival analysis: A self-learning text*. Springer, 1996.

Klimt, B. and Yang, Y. Introducing the Enron Corpus. In *Proceedings of the First Conference on Email and Anti-Spam (CEAS)*, Mountain View, CA, 2004. URL http://www.ceas.cc/papers-2004/168.pdf.

Knoth, S. and Knoth, S. Exact average run lengths of cusum schemes for erlang distributions: Exact average run lengths of cusum schemes. *Sequential Analysis*, 17 (2):173–184, 1998.

Kwapisz, J. R., Weiss, G. M., and Moore, S. A. Activity recognition using cell phone accelerometers. *ACM SigKDD Explorations Newsletter*, 12(2):74–82, 2011.

Lachiche, N., Ferri, C., Macskassy, S., and Rakotomamonjy, A. Proceedings of the second workshop on ROC analysis in machine learning (ROCML'05), 2005. URL https://publis.icube.unistra.fr/index.php/11-LFMR05.

Lachiche, N., Ferri, C., and Macskassy, S. Proceedings of the third workshop on ROC analysis in machine learning (ROCML'06), 2006. URL https://dmip.webs.upv.es/ROCML2006/.

Lavielle, M. Detection of multiple changes in a sequence of dependent variables. *Stochastic Processes and their applications*, 83(1):79–102, 1999.

Lavielle, M. and Teyssiere, G. Detection of multiple changepoints in multivariate time series. *Lithuanian Mathematical Journal*, 46(3):287–306, 2006.

Lavin, A. and Ahmad, S. Evaluating real-time anomaly detection algorithms – the numenta anomaly benchmark. In *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*, pp. 38–44, 2015. doi: 10.1109/ICMLA.2015.141.

Li, S., Cao, Y., Leamon, C., Xie, Y., Shi, L., and Song, W. Online seismic event picking via sequential change-point detection. In *2016 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 774–779. IEEE, 2016.

Li, Z., Zou, C., Gong, Z., and Wang, Z. The computation of average run length and average time to signal: an overview. *Journal of Statistical Computation and Simulation*, 84(8):1779–1802, 2014.

Lim, J. and Lee, S. Efficient ARL estimation for general control charts using censored run lengths. *Quality Engineering*, 37(3):359–368, 2025.

Lin, H.-M., Liu, S. T., Levin, M. A., Williamson, J., Bouvier, N. M., Aberg, J. A., Reich, D., and Egorova, N. Informative censoring—a cause of bias in estimating COVID-19 mortality using hospital data. *Life*, 13(1):210, 2023.

Lorden, G. Procedures for reacting to a change in distribution. *The annals of mathematical statistics*, pp. 1897–1908, 1971.

Makone, A. Twitter US Airline Sentiment, 2016. URL https://www.kaggle.com/datasets/crowdflower/twitter-airline-sentiment. Original data from CrowdFlower's "Data for Everyone" library; accessed 2025-11-24.

Malmquist, S. What rates of censoring induce bias in the Kaplan–Meier estimator for small samples? Master's thesis, Uppsala University, Department of Statistics, 2025.

Mei, Y. Is average run length to false alarm always an informative criterion? *Sequential Analysis*, 27(4):354–376, 2008.

Mills, M. Competing risk and multi-state models. *Introducing Survival and Event History Analysis*, pp. 190–212, 2011.

Mishra, J. and Prasanna, S. M. Spoken language change detection inspired by speaker change detection. *Circuits, Systems, and Signal Processing*, 43(10):6373–6398, 2024.

Morita, K. Introduction to survival analysis in the presence of competing risks. *Annals of Clinical Epidemiology*, 3 (4):97–100, 2021.

Mousavi, S. M., Sheng, Y., Zhu, W., and Beroza, G. C. Stanford earthquake dataset (STEAD): A global data set of seismic signals for AI. *IEEE Access*, 7:179464–179476, 2019.

Niu, Y. S., Hao, N., and Zhang, H. Multiple change-point detection: A selective overview. *Statistical Science*, pp. 611–623, 2016.

Page, E. S. Continuous inspection schemes. *Biometrika*, 41(1/2):100–115, 1954. ISSN 00063444. URL http://www.jstor.org/stable/2333009.

Pardo-Fernández, J. C. and Castro, P. R. International workshop on ROC analysis and related topics (ROC2025): Programme and book of abstracts, January 2025. URL https://roc2025.webs.uvigo.es/programme_and_BoA_ROC2025.pdf. International Workshop on ROC Analysis and Related Topics, Vigo, January 23–24, 2025.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. Curran Associates, Inc., 2019. License: https://github.com/pytorch/pytorch/blob/master/LICENSE.

Pehlivan, C. and Testik, M. C. Impact of model misspecification on the exponential EWMA charts: a robustness study when the time-between-events are not exponential. *Quality and Reliability Engineering International*, 26(2): 177–190, 2010.

Phinikettos, I. and Gandy, A. An omnibus CUSUM chart for monitoring time to event data. *Lifetime data analysis*, 20(3):481–494, 2014.

Pollak, M. Optimal detection of a change in distribution. *The Annals of Statistics*, pp. 206–227, 1985.

Polunchenko, A. S. Exact distribution of the generalized Shiryaev–Roberts stopping time under the minimax Brownian motion setup. *Sequential Analysis*, 35(1): 108–143, 2016. doi: 10.1080/07474946.2016.1132066. URL https://www.tandfonline.com/doi/abs/10.1080/07474946.2016.1132066.

Poor, H. V. Quickest detection with exponential penalty for delay. *The Annals of Statistics*, 26(6):2179–2205, 1998.

Purohit, H., Tanabe, R., Ichige, K., Endo, T., Nikaido, Y., Suefusa, K., and Kawaguchi, Y. MIMII dataset: Sound dataset for malfunctioning industrial machine investigation and inspection. *arXiv preprint arXiv:1909.09347*, 2019.

Qi, T. and Wang, J. Calculating restricted mean survival time. In *PharmaSUG 2018*, 2018. URL https://pharmasug.org/proceedings/2018/AA/PharmaSUG-2018-AA04.pdf.

Qiu, P. *Introduction to statistical process control*. CRC press, 2013.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021. URL https://www.R-project.org/.

Ranganathan, P. and Pramesh, C. Censoring in survival analysis: potential for bias. *Perspectives in clinical research*, 3(1):40, 2012.

Reynolds, M. R. Approximations to the average run length in cumulative sum control charts. *Technometrics*, 17(1): 65–71, 1975.

Rivest, L.-P. and Wells, M. T. A martingale approach to the copula-graphic estimator for the survival function under dependent censoring. *Journal of Multivariate Analysis*, 79(1):138–155, 2001.

Roberts, S. W. Control chart tests based on geometric moving averages. *Technometrics*, 1(3):239–250, 1959. ISSN 00401706. URL http://www.jstor.org/stable/1266443.

Romano, G., Eckley, I. A., and Fearnhead, P. A log-linear nonparametric online changepoint detection algorithm based on functional pruning. *IEEE Transactions on Signal Processing*, 72:594–606, 2024a. doi: 10.1109/TSP.2023.3343550.

Romano, G., Grose, D., Ward, K., Edward, A., Pishchagina, L., Rigaill, G., Runge, V., Fearnhead, P., and Eckley, I. A. changepoint_online: A collection of methods for online changepoint detection, April 2024b. URL https://github.com/grosed/changepoint_online.

Royston, P. and Parmar, M. K. Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. *BMC medical research methodology*, 13(1): 152, 2013.

Ryant, N., Church, K., Cieri, C., Cristia, A., Du, J., Ganapathy, S., and Liberman, M. First DIHARD challenge evaluation plan. In *Technical Report*. Linguistic Data Consortium, University of Pennsylvania, 2018.

Saccucci, M. S. and Lucas, J. M. Average run lengths for exponentially weighted moving average control schemes using the Markov chain approach. *Journal of Quality Technology*, 22(2):154–162, 1990.

Sahki, N., Gégout-Petit, A., and Wantz-Mézières, S. Performance study of change-point detection thresholds for cumulative sum statistic in a sequential context. *Quality and Reliability Engineering International*, 36(8):2699–2719, 2020.

Salles, R., Lima, J., Reis, M., Coutinho, R., Pacitti, E., Masseglia, F., Akbarinia, R., Chen, C., Garibaldi, J., Porto, F., et al. SoftED: Metrics for soft evaluation of time series event detection. *Computers & Industrial Engineering*, 198:110728, 2024.

Sasikumr, R. and Sujatha, M. A comprehensive study on monitoring treatment outcomes using risk-adjusted CUSUM control charts. *Journal of Information Systems Engineering and Management*, 10(35s):248–263, April 2025. ISSN 2468-4376. doi: 10.52783/jisem.v10i35s.5988. URL https://jisem-journal.com/index.php/journal/article/view/5988.

Sego, L. H., Reynolds Jr, M. R., and Woodall, W. H. Risk-adjusted monitoring of survival times. *Statistics in medicine*, 28(9):1386–1401, 2009.

Shafiq, M., Shah, S., and Alamgir, M. Modified weighted Kaplan-Meier estimator. *Pakistan Journal of Statistics and Operation Research*, pp. 39–44, 2007.

Shiryaev, A. N. On optimum methods in quickest detection problems. *Theory of Probability & Its Applications*, 8(1): 22–46, 1963.

Stolfo, S., Fan, W., Lee, W., Prodromidis, A., and Chan, P. KDD Cup 1999 Data. UCI Machine Learning Repository, 1999. DOI: https://doi.org/10.24432/C51C7N.

Stute, W. The bias of Kaplan-Meier integrals. *Scandinavian Journal of Statistics*, pp. 475–484, 1994.

Stute, W. and Wang, J.-L. The strong law under random censorship. *The Annals of statistics*, pp. 1591–1607, 1993.

Sunthornwat, R., Sukparungsee, S., and Areepong, Y. Analytical explicit formulas of average run length of homogenously weighted moving average control chart based on a MAX process. *Symmetry*, 15(12):2112, 2023.

Tan, E., Algar, S. D., Corrêa, D., Stemler, T., and Small, M. Network representations of attractors for change point detection. *Communications Physics*, 6(1):340, 2023.

Tartakovsky, A. *Sequential change detection and hypothesis testing: General non-iid stochastic models and asymptotically optimal rules*. Chapman and Hall/CRC, 2019.

Tartakovsky, A., Nikiforov, I., and Basseville, M. *Sequential Analysis: Hypothesis Testing and Changepoint Detection*. Chapman & Hall/CRC, 1st edition, 2014.

Tatbul, N., Lee, T. J., Zdonik, S., Alam, M., and Gottschlich, J. Precision and recall for time series. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/8f468c873a32bb0619eaeb2050ba45d1-Paper.pdf.

Therneau, T., Crowson, C., and Atkinson, E. Multi-state models and competing risks. *CRAN-R (https://cran.r-project.org/web/packages/survival/vignettes/compete.pdf)*, 2020.

Truong, C., Oudre, L., and Vayatis, N. Selective review of offline change point detection methods. *Signal Processing*, 167:107299, 2020.

Tutz, G., Schmid, M., et al. *Modeling discrete time-to-event data*, volume 3. Springer, 2016.

Van den Burg, G. J. J. and Williams, C. K. I. An evaluation of change point detection algorithms. *arXiv preprint arXiv:2003.06222*, 2020.

Van Rossum, G. and Drake, F. L. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA, 2009. ISBN 1441412697.

Wadinger, M., Kvasnica, M., and Kawahara, Y. Changepoint detection in industrial data streams based on online dynamic mode decomposition with control. *arXiv preprint arXiv:2407.05976*, 2024.

Wang, B., Zhao, Y., Yang, L., Long, T., and Li, X. Temporal action localization in the deep learning era: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(4):2171–2190, 2023a.

Wang, X., Borsoi, R. A., Richard, C., and Chen, J. Change point detection with neural online density-ratio estimator. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023b.

Woollam, J., Münchmeyer, J., Tilmann, F., Rietbrock, A., Lange, D., Bornstein, T., Diehl, T., Giunchi, C., Haslinger, F., Jozinović, D., et al. SeisBench—a toolbox for machine learning in seismology. *Seismological Society of America*, 93(3):1695–1709, 2022.

Xiang, Y., Akcakaya, M., Sen, S., and Nehorai, A. Target detection via cognitive radars using change-point detection, learning, and adaptation. *Circuits, Systems, and Signal Processing*, 40(1):233–261, 2021.

Zare, A. and Mahmoodi, M. Modified Kaplan-Meier estimator based on competing risks for heavy censoring data. *Int J Statist Med Res*, 2(4):297–304, 2013.

Appendix

A. QCD-Survival Analysis Correspondence

Proposed Correspondence. We summarize our key idea, an analogy between QCD and survival analysis in Tab. 1.

ARL	ADD	Survival analysis
detection time τ_i	detection delay $\tau_i - \nu_i$	event time
$\min\{\nu_i, T_i\}$	$T_i - \nu_i$	censoring time C_i
KM-ARL $\hat{\mu}_T^{(\mathrm{KM})}$	KM-ADD $\hat{M}_T^{(\mathrm{KM})}$	restricted mean survival time

Table 1. QCD-survival analysis analogy.

Illustration. For clarity, we illustrate the original KME, KM-ARL, and the computation of d_j and n_j in Fig. 5.

A.1. Background of KME in Survival Analysis

A.1.1. Introduction

We provide the motivation and intuition for the original KME in survival analysis (Kaplan & Meier, 1958; Kleinbaum & Klein, 1996). Let τ be a nonnegative random variable representing a lifetime, and let

$$S(t) := P(\tau > t) \tag{12}$$

denote the survival function. In the ideal (textbook) setting without censoring, and with i.i.d. samples $\tau_1, \ldots, \tau_N$, a naive nonparametric estimator of $S(t)$ is the following empirical survival function

$$\hat{S}_{\mathrm{emp}}(t) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\{\tau_i > t\}. \tag{13}$$

However, in practice, we often do *not* observe all τ_i fully. Instead, some subjects drop out, are lost to follow-up, or the study ends before they fail. This leads to *right-censoring*, where we observe

$$Y_i = \min(\tau_i, C_i), \qquad \Delta_i = \mathbb{1}\{\tau_i \le C_i\}, \tag{14}$$

with C_i a censoring time. Then Y_i is either the failure time (if $\Delta_i = 1$) or a censoring time (if $\Delta_i = 0$), where Δ_i is the event flag.

If we treat censored observations as if they were failures, we underestimate survival; if we drop them entirely, we discard information about the period in which we do know they survived. The KME is motivated precisely by this need:

- use all information available prior to censoring;
- do not make parametric assumptions (nonparametric);
- remain interpretable as a product of empirical conditional survival probabilities.

We derive the KME formula in the following to clarify its motivation.

A.1.2. Decomposition via Conditional Survival Probabilities

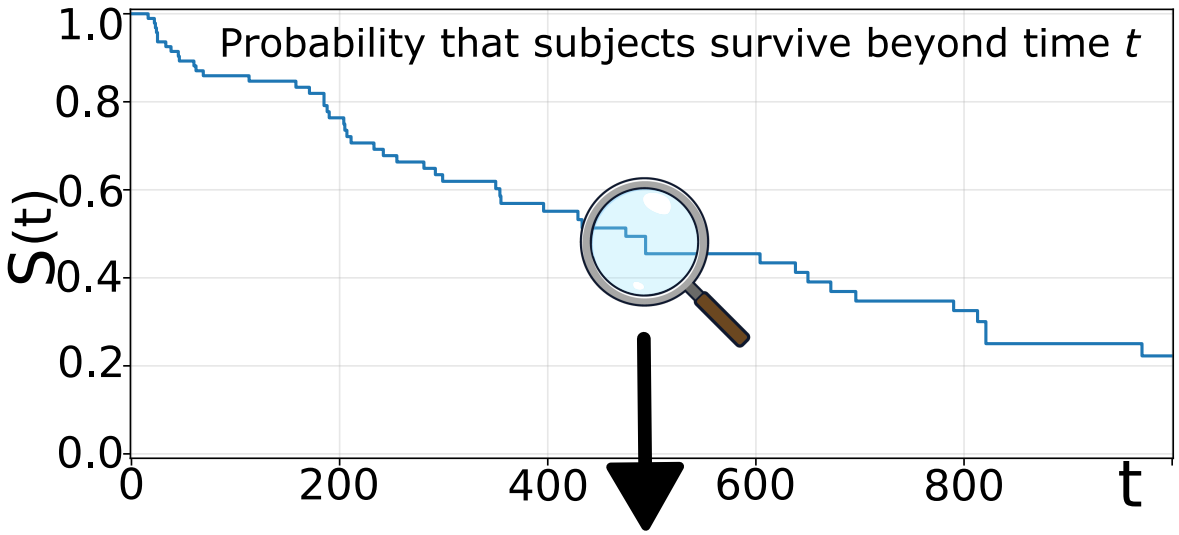
A key idea is to write $S(t)$ as a product of conditional survival probabilities at each distinct failure time (or event time, i.e., the time of death). Let

$$t_{(1)} < t_{(2)} < \cdots < t_{(K)} \tag{15}$$

be the distinct *failure* times in the population (not yet the sample). For t between $t_{(j)}$ and $t_{(j+1)}$, we can write

$$S(t) = P(\tau > t) = \prod_{t_{(m)} \le t} P\big(\tau > t_{(m)} \,\big|\, \tau \ge t_{(m)}\big). \tag{16}$$

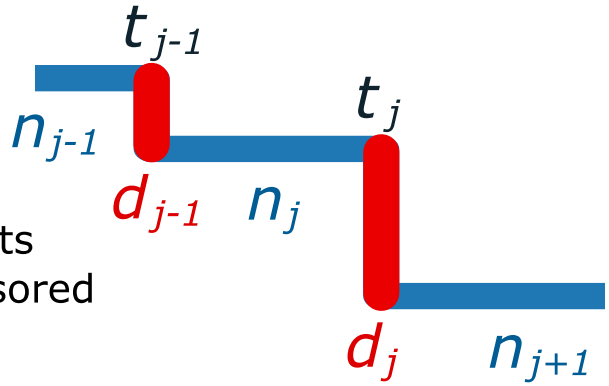
This is just repeated application of the chain rule of probability. Intuitively,



Survival Analysis

d_j the number of subjects died at t_j

n_j the number of subjects neither dead nor censored prior to t_j



QCD

d_j^{ARL} the number of sequences with detection at t_j^{ARL}

n_j^{ARL} the number of sequences neither detected nor truncated prior to t_j^{ARL}

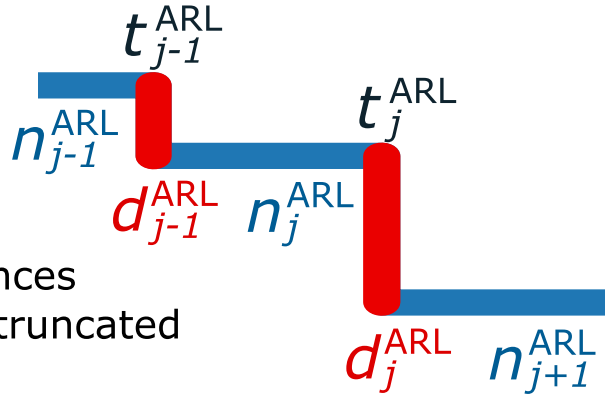


Figure 5. **Original KME, KM-ARL, and computation of d_j and n_j.** we estimate the *survival function of detection points* $S^{\mathrm{ARL}}(t) := P(\tau > t \mid \nu = \infty)$ using the Kaplan-Meier estimator (KME) (Kaplan & Meier, 1958), a non-parametric estimator of the survival function: $\hat{S}^{\mathrm{ARL}}(t) = \prod_{j:t_j^{\mathrm{ARL}} \le t}(1 - \frac{d_j^{\mathrm{ARL}}}{n_j^{\mathrm{ARL}}})$, where $0 < t_1^{\mathrm{ARL}} < t_2^{\mathrm{ARL}} < \cdots < t_{N'}^{\mathrm{ARL}}$ are distinct detection points, with $N'_{\mathrm{ARL}} \in \mathbb{N}$ $(\le N)$; $d_j^{\mathrm{ARL}} := |\{i \in [N] \mid \tau_i = t_j^{\mathrm{ARL}}\}|$ is the number of sequences with a detection at t_j^{ARL} $(j \in [N'_{\mathrm{ARL}}])$; and $n_j^{\mathrm{ARL}} := |\{i \in [N] \mid \min\{\tau_i, C_i^{\mathrm{ARL}}\} \ge t_j^{\mathrm{ARL}}\}|$ is the number of sequences neither detected nor censored prior to t_j^{ARL} $(j \in [N'_{\mathrm{ARL}}])$, with the censoring time of sequence i defined as $C_i^{\mathrm{ARL}} := \min\{\nu_i, T_i\}$. We propose a non-parametric estimator of the ARL under irregular sequence lengths, termed KM-ARL, as the integral of $\hat{S}^{\mathrm{ARL}}(t)$ over the range $[0, a]$ for arbitrary $a \in \mathbb{R}_{\ge 0}$: $\hat{\mu}_T^{(\mathrm{KM})} := \int_0^a \hat{S}^{\mathrm{ARL}}(t)dt$.

- at each time $t_{(m)}$, we look at those who are still alive (have "survived so far");
- we multiply by the probability that they survive *beyond* $t_{(m)}$;
- the overall survival up to time t is the product of these step-by-step survival probabilities.

The KM estimator simply replaces each conditional probability $P(\tau > t_{(m)} \mid \tau \geq t_{(m)})$ by a natural empirical estimate using the sample with censoring.

A.1.3. Risk Set n_j and Number of Events d_j

Suppose we have N subjects, and we observe (Y_i, Δ_i) for $i = 1, \ldots, N$. Let

- $t_{(1)} < \cdots < t_{(J)}$ be the distinct observed *failure* times in the sample (i.e., times where at least one $\Delta_i = 1$ and $Y_i = t_{(j)}$),
- d_j be the number of failures at time $t_{(j)}$,
- n_j be the size of the *risk set* at time $t_{(j)}$, i.e., the number of subjects who are known to be alive and uncensored just *before* $t_{(j)}$.

Intuition for the risk set n_j. The risk set n_j at time $t_{(j)}$ collects all individuals for whom failure at $t_{(j)}$ is still a possibility:

- subjects who failed earlier ($Y_i < t_{(j)}$ and $\Delta_i = 1$) are removed: they are already dead;
- subjects who were censored earlier ($Y_i < t_{(j)}$ and $\Delta_i = 0$) are removed: we no longer observe them, and we cannot know whether they would have failed at $t_{(j)}$ or not;
- subjects whose observed time Y_i is at least $t_{(j)}$ remain in the risk set: we know they survived at least up to just before $t_{(j)}$.

Thus,

$$n_j = \sum_{i=1}^{N} \mathbb{1}\{Y_i \geq t_{(j)}\}. \tag{17}$$

This definition uses all partial information: even if some Y_i correspond to future censoring events, until they are censored, they contribute to the information that the subject has survived up to that time.

Intuition for the number of failures d_j. At $t_{(j)}$, we also count the number of failures:

$$d_j = \sum_{i=1}^{N} \mathbb{1}\{Y_i = t_{(j)},\ \Delta_i = 1\}. \tag{18}$$

Only *failures* contribute to d_j. Censoring at time $t_{(j)}$ does not count as a failure; it simply removes subjects from future risk sets (for times after $t_{(j)}$). This aligns with the goal of estimating the distribution of failure times, not censoring times.

A.1.4. Why d_j/n_j? Conditional Probability Viewpoint

Consider the underlying quantity

$$P(\tau = t_{(j)} \mid \tau \geq t_{(j)}) \tag{19}$$

as the probability that a subject who has survived up to $t_{(j)}$ fails exactly at $t_{(j)}$. In a discrete-time picture, this is analogous to a "hazard probability" at $t_{(j)}$. Given n_j subjects in the risk set at $t_{(j)}$, and d_j of them failing at $t_{(j)}$, it is natural to estimate this conditional probability by the empirical proportion

$$\hat{P}(\tau = t_{(j)} \mid \tau \geq t_{(j)}) = \frac{d_j}{n_j}. \tag{20}$$

Consequently, the conditional survival probability at $t_{(j)}$ is estimated by

$$\hat{P}(\tau > t_{(j)} \mid \tau \geq t_{(j)}) = 1 - \frac{d_j}{n_j}. \tag{21}$$

Note that censored individuals are *included* in n_j as long as they are not yet censored at $t_{(j)}$, but never in d_j. This reflects the idea that:

- up to the censoring time, their "survival experience" is fully observed and contributes to the risk set n_j;
- at the censoring time, we stop knowing what happens afterward, so they drop from future risk sets (but are not treated as failures d_j).

A.1.5. KME AS PRODUCT-LIMIT ESTIMATOR

Putting everything together, the KM estimator of $S(t)$ is defined as

$$\hat{S}_{\mathrm{KM}}(t) = \prod_{t_{(j)} \leq t} \left(1 - \frac{d_j}{n_j}\right). \tag{22}$$

This is exactly the product of estimated conditional survival probabilities:

$$\hat{S}_{\mathrm{KM}}(t) = \prod_{t_{(j)} \leq t} \hat{P}(\tau > t_{(j)} \mid \tau \geq t_{(j)}). \tag{23}$$

Intuitively:

- At the beginning ($t < t_{(1)}$), survival is 1: no one has failed.
- At each failure time $t_{(j)}$, we shrink the survival curve by multiplying by $1 - d_j/n_j$.
- Between failure times, $\hat{S}_{\mathrm{KM}}(t)$ is constant, yielding a right-continuous step function.

This construction uses only the ordering of failure and censoring times and assumes (in the standard theory) that censoring is non-informative (independent of the failure mechanism, given the observed history). Under these conditions, the KM estimator can be seen as the nonparametric maximum likelihood estimator of $S(t)$ under right-censoring.

In summary, the roles of d_j and n_j are:

- n_j is the size of the risk set just before $t_{(j)}$. It collects all individuals whose failure at $t_{(j)}$ is still observable: they have not failed and have not been censored yet. This ensures that d_j/n_j is a valid empirical conditional probability.
- d_j is the number of failures at $t_{(j)}$. It quantifies how much the survival curve should drop at that time, relative to the current risk set.
- The ratio d_j/n_j is the natural empirical estimator of the conditional failure probability at $t_{(j)}$. Its complement $1-d_j/n_j$ is the empirical conditional survival probability.
- The KM estimator is then the product over time of these conditional survival probabilities, giving a nonparametric estimator of the entire survival function that properly accounts for right-censoring.

B. Proofs

Our proofs for the finite-sample bounds are based on (Stute, 1994). In our proofs for the finite-sample bounds, we idealize the time index as continuous without loss of generality, for technical convenience. Define $\nu_1, \ldots, \nu_N$ as i.i.d. random variables (the changepoints), independent of the observations of sequence, where $N \in \mathbb{N}$. Assume that $\tau_1, \ldots, \tau_N$ are independent and identically distributed (i.i.d.) random variables (the detection points of a QCD model). Assume that $T_1, \ldots, T_N$ are i.i.d. random variables (the lengths of sequences), independent of the observations of sequences.

B.1. Proof of Theorem 4.1

Let $C_1^{\mathrm{ARL}}, \ldots, C_N^{\mathrm{ARL}}$ are i.i.d. random variables (the censoring times for the KM-ARL), defined as $C_i^{\mathrm{ARL}} := \min\{\nu_i, T_i\}$ for $i \in [N]$. Assume that C^{ARL} is independent of τ, called the independent censoring assumption (or non-informative censoring assumption) for the KM-ARL. This assumption holds for online QCD models that do not look ahead the input sequence, which are the focus of this paper. This is because: detection points τ_i prior to censoring can be regarded as samples from $P(\tau \mid \nu = \infty)$; $P(\tau \mid \nu = \infty)$, $P(\nu)$, and $P(T)$ are independent; and thus, $P(\tau \mid \nu = \infty)$ and $P(C^{\mathrm{ARL}} := \min\{\nu, T\})$ are also independent. Here, P denotes a probability or a probability measure, with slight abuse of notation throughout the paper. In contrast, the independent censoring assumption does not hold for offline changepoint detection models because τ depends on $X^{(0,T)}$, which in turn depends on ν and T. Let F^{ARL} and G^{ARL} denote the cumulative distribution functions (CDFs) of τ and C^{ARL}, respectively. The survival function of detection points is defined as $S^{\mathrm{ARL}}(t) := 1 - F^{\mathrm{ARL}}(t)$, where $t \in \mathbb{R}_{\geq 0}$.

We adopt a different convention of the KME from that in the main text. Consider the random variables

$$Z_i^{\mathrm{ARL}} := \min(\tau_i, C_i^{\mathrm{ARL}}), \tag{24}$$

$$\delta_i^{\mathrm{ARL}} := \mathbb{1}_{\{\tau_i < C_i^{\mathrm{ARL}}\}}, \tag{25}$$

where δ_i^{ARL} is the indicator representing whether detection τ_i has been observed before a changepoint or the end of the sequence. Let H^{ARL} denote the CDF of Z^{ARL}, and assume that it is continuous on $t \in \mathbb{R}_{\geq 0}$. Let $Z_{1:N}^{\mathrm{ARL}} \leq Z_{2:N}^{\mathrm{ARL}} \leq \cdots \leq Z_{N:N}^{\mathrm{ARL}}$ be the *order statistics* of Z^{ARL} (i.e., Z_i^{ARL} are sorted in increasing order) and $\delta_{[i:N]}^{\mathrm{ARL}}$ be the *concomitant* of $Z_{i:N}^{\mathrm{ARL}}$; i.e., δ_i^{ARL} are sorted according to $Z_{i:N}^{\mathrm{ARL}}$, meaning that $\delta_{[i:N]}^{\mathrm{ARL}} = \delta_j$ if $Z_{i:N}^{\mathrm{ARL}} = Z_j^{\mathrm{ARL}}$. The estimation of $S^{\mathrm{ARL}}(t)$ is given by

$$1 - \hat{F}_N^{\mathrm{ARL}}(t) := \hat{S}^{\mathrm{ARL}}(t) = \prod_{i=1}^{N} \left(1 - \frac{\delta_{[i:N]}^{\mathrm{ARL}}}{N - i + 1}\right)^{\mathbb{1}(Z_{i:N}^{\mathrm{ARL}} \leq t)}, \tag{26}$$

where $t \in \mathbb{R}_{\geq 0}$ This is equivalent to the convention in the main text, which can be confirmed by canceling factors telescopically. In case of ties, treat a death as if it occurs before a censoring at the same time point. Our KM-ARL is formally defined as

$$\hat{\mu}_T^{(\mathrm{KM})} = \int_0^a t \, d\hat{F}_N^{\mathrm{ARL}}, \tag{27}$$

where $a \in \mathbb{R}_{\geq 0}$ is arbitrary.[1]

Theorem B.1 (Finite-sample bias bounds for KM-ARL (Thm. 4.1))**.** *We idealize the time index as continuous without loss of generality, for technical convenience. Let $F^{\mathrm{ARL}}, G^{\mathrm{ARL}}$, and H^{ARL} be the CDFs of τ, C^{ARL}, and Z^{ARL}, respectively. Assume that (i) F^{ARL} and G^{ARL} do not have common discontinuities, (ii) τ and C^{ARL} are independent, known as the independent censoring (or non-informative censoring), and (iii) H^{ARL} is continuous. Then, we have for any $a \in \mathbb{R}_{\geq 0}$*

$$-\int_0^a t \, G^{\mathrm{ARL}}(t) \, H^{\mathrm{ARL}}(t)^{N-1} \, F^{\mathrm{ARL}}(dt) \tag{28}$$

$$\leq \mathcal{B}_{\mathrm{FS}}(\hat{\mu}_T^{(\mathrm{KM})}) \leq \tag{29}$$

$$\int_0^a a \, G^{\mathrm{ARL}}(t) \, H^{\mathrm{ARL}}(t)^{N-1} \, F^{\mathrm{ARL}}(dt), \tag{30}$$

Assumption (i) is a technical requirement and is valid unless pathological situations are considered. In most cases, it holds with probability 1 because we consider continuous time. Note that F^{ARL} and G^{ARL} may have separate discontinuities. Assumption (ii) holds for online QCD models that do not look ahead the input sequence, which are the focus of this paper. This is because: detection points τ_i prior to censoring can be regarded as samples from $P(\tau \mid \nu = \infty)$; $P(\tau \mid \nu = \infty)$, $P(\nu)$, and $P(T)$ are independent; and thus, $P(\tau \mid \nu = \infty)$ and $P(C := \min\{\nu, T\})$ are also independent. In contrast, Assumption (ii) does not hold for offline changepoint detection models because τ depends on $X^{(0,T)}$, which in turn depends on ν and T. Assumption (iii) is also a technical requirement.

[1] This definition can be rephrased as Eq. (3) because $\int_0^a S(t) = \int_0^a tF(t)dt + aS(a)$. The gap is given by a constant $aS(a)$.

Proof.

$$\begin{aligned}
&\int_0^a t\, dF^{\mathrm{ARL}}(t) \\
&= \int_0^a \Big(\int_0^a \mathbb{1}(t > s)\, ds \Big) dF^{\mathrm{ARL}}(t) && (31)\\
&= \int_0^a \Big(\int_0^a \mathbb{1}(t > s)\, dF^{\mathrm{ARL}}(t) \Big) ds && (32)\\
&= \int_0^a \Big(\int_s^a dF^{\mathrm{ARL}}(t) \Big) ds && (33)\\
&= \int_0^a \big(F^{\mathrm{ARL}}(a) - F^{\mathrm{ARL}}(s) \big) ds && (34)\\
&= aF^{\mathrm{ARL}}(a) - a + \int_0^a S^{\mathrm{ARL}}(t)\, dt. && (35)
\end{aligned}$$

Similarly, we have

$$\int_0^a t\, d\hat{F}^{\mathrm{ARL}}(t) = a\hat{F}^{\mathrm{ARL}}(a) - a + \int_0^a \hat{S}^{\mathrm{ARL}}(t)\, dt. \tag{36}$$

Thus, the finite-sample bias is given by

$$\begin{aligned}
&B_{\mathrm{FS}}\big(\hat{\mu}_T^{(\mathrm{KM})}\big) \\
=&\mathbb{E}\Big[\int_0^a \hat{S}^{\mathrm{ARL}}(t)\, dt\Big] - \int_0^a S^{\mathrm{ARL}}(t)\, dt \\
=&\mathbb{E}\Big[\int_0^a t\, d\hat{F}^{\mathrm{ARL}}(t)\Big] - \int_0^a t\, dF^{\mathrm{ARL}}(t) \\
&- a\big(\mathbb{E}[\hat{F}^{\mathrm{ARL}}(a)] - F^{\mathrm{ARL}}(a)\big) \\
=:&I_1 - aI_2, && (37)
\end{aligned}$$

where we defined

$$I_1 = \mathbb{E}\Big[\int_0^a t\, d\hat{F}^{\mathrm{ARL}}(t)\Big] - \int_0^a t\, dF^{\mathrm{ARL}}(t) \tag{38}$$

$$I_2 = \mathbb{E}[\hat{F}^{\mathrm{ARL}}(a)] - F^{\mathrm{ARL}}(a). \tag{39}$$

We bound them by invoking the following lemma from (Stute, 1994), adapted to our setting. The full proof of Lem. B.2 is lengthy but given in (Stute, 1994). Refer also to (Stute & Wang, 1993), on which Stute (1994) builds his result.

Lemma B.2 (Cor. 1.1 in (Stute, 1994))**.** *For any F^{ARL}, G^{ARL}, H^{ARL}, and $\hat{F}^{\mathrm{ARL}}$ that satisfy the assumptions described in this section, the following inequality holds for any $N \in \mathbb{N}$, where $\varphi(t) \geq 0$ is a Borel measurable function on $t \in \mathbb{R}_{\geq 0}$:*

$$\begin{aligned}
&-\int \varphi(t)\, G^{\mathrm{ARL}}(t)\, H^{\mathrm{ARL}}(t)^{N-1}\, dF^{\mathrm{ARL}}(t) \\
&\leq \mathbb{E}[\int \varphi(t) d\hat{F}^{\mathrm{ARL}}(t)] - \int \varphi(t) dF^{\mathrm{ARL}} \\
&\leq 0. && (40)
\end{aligned}$$

According to Lem. B.2, for $\varphi(t) = \mathbb{1}(t \in [0, a])$, we have

$$\begin{aligned}
&-\int_0^a G^{\mathrm{ARL}}(t)\, H^{\mathrm{ARL}}(t)^{N-1}\, dF^{\mathrm{ARL}}(t) \\
&\leq \mathbb{E}[\int_0^a d\hat{F}^{\mathrm{ARL}}(t)] - \int_0^a dF^{\mathrm{ARL}} \leq 0. && (41)
\end{aligned}$$

Since $\mathbb{E}[\int_0^a d\hat{F}^{\mathrm{ARL}}(t)] - \int_0^a dF^{\mathrm{ARL}} = \mathbb{E}[\hat{F}^{\mathrm{ARL}}(a)] - F^{\mathrm{ARL}}(a) = I_2$, we have

$$-\int_0^a G^{\mathrm{ARL}}(t)\, H^{\mathrm{ARL}}(t)^{N-1}\, dF^{\mathrm{ARL}}(t) \leq I_2 \leq 0. \tag{42}$$

Additionally, for $\varphi(t) = t\,\mathbb{1}(t \in [0, a])$, we have

$$\begin{aligned} &-\int_0^a t\, G^{\mathrm{ARL}}(t)\, H^{\mathrm{ARL}}(t)^{N-1}\, dF^{\mathrm{ARL}}(t) \\ &\leq \mathbb{E}[\int_0^a t d\hat{F}^{\mathrm{ARL}}(t)] - \int_0^a t dF^{\mathrm{ARL}}(= I_1) \leq 0. \end{aligned} \tag{43}$$

From Eqs. (37), (38), (39), (42), and (43), we have proved

$$\begin{aligned} &-\int_0^a t\, G^{\mathrm{ARL}}(t)\, H^{\mathrm{ARL}}(t)^{N-1}\, dF^{\mathrm{ARL}}(t) \\ &\leq I_1 - aI_2 \;\; (= B_{\mathrm{FS}}(\hat{\mu}_T^{(\mathrm{KM})})) \leq \\ &\int_0^a a\, G^{\mathrm{ARL}}(t)\, H^{\mathrm{ARL}}(t)^{N-1}\, dF^{\mathrm{ARL}}(t) \end{aligned} \tag{44}$$

for arbitrary $a \in \mathbb{R}_{\geq 0}$. □

B.2. Proof of Theorem 4.2

Consider only sequences with $\nu < \infty$ and $\Delta\tau \geq 0$ only, as others do not contribute to the computation of the ADD, defined as $\mathbb{E}[\tau - \nu \mid \tau \geq \nu, \nu < \infty]$. In this section, we use $N \in \mathbb{N}$ as the number of sequences with $\nu_i < \infty$ and $\Delta\tau_i \geq 0$ in the dataset. Assume that $\Delta\tau_1, \ldots, \Delta\tau_N$ are independent and identically distributed (i.i.d.) random variables (the detection delays of a QCD model), defined as $\Delta\tau_i := \tau_i - \nu_i$. Let $C_1^{\mathrm{ADD}}, \ldots, C_N^{\mathrm{ADD}}$ are i.i.d. random variables (the censoring times for the KM-ADD), defined as $C_i^{\mathrm{ADD}} := \Delta T_i = T_i - \nu_i$ for $i \in [N]$. Assume that C^{ADD} is independent of $\Delta\tau$, called the independent censoring assumption (or non-informative censoring assumption) for the KM-ADD. The independent censoring assumption is justified in online QCD models by the approximate $P(\Delta\tau \mid \Delta\tau \geq 0, \nu < \infty) \approx P(\tau \mid \nu = 0)$ and $P(C := \Delta T \mid \Delta\tau \geq 0, \nu < \infty) \approx P(C := T \mid \nu = 0)$ because $P(\tau \mid \nu = 0)$ and $P(C := T \mid \nu = 0)$ are independent. This indicates that the distributions of detection delay $\tau - \nu$ and censoring time $T - \nu$ are approximately equal to the distributions of them measured from $t = 0$, not ν. Under this approximation, the definition of the ADD becomes $\mathbb{E}[\tau \mid \nu = 0]$, which is called the steady-state ARL (Saccucci & Lucas, 1990; Sasikumr & Sujatha, 2025; Lim & Lee, 2025) and often used in the theory of control charts. For simplicity, we assume the independent censoring in the proof of Thm. 4.2, and define F^{ADD}, G^{ADD}, and H^{ADD} accordingly. The survival function of detection delays is defined as $S^{\mathrm{ADD}}(t) := 1 - F^{\mathrm{ADD}}(t)$, where $t \in \mathbb{R}_{\geq 0}$.

Consider the random variables

$$Z_i^{\mathrm{ADD}} := \min(\Delta\tau_i, C_i^{\mathrm{ADD}}), \tag{45}$$

$$\delta_i^{\mathrm{ADD}} := \mathbb{1}_{\{\Delta\tau_i < C_i^{\mathrm{ADD}}\}}, \tag{46}$$

where δ_i^{ADD} is the indicator representing whether the detection τ_i has been observed before the end of the sequence. Let H^{ADD} denote the CDF of Z^{ADD}, and assume that it is continuous on $t \in \mathbb{R}_{\geq 0}$. Let $Z_{1:N}^{\mathrm{ADD}} \leq Z_{2:N}^{\mathrm{ADD}} \leq \cdots \leq Z_{N:N}^{\mathrm{ADD}}$ be the *order statistics* of Z^{ADD} (i.e., Z_i^{ADD} are sorted in increasing order) and $\delta_{[i:N]}^{\mathrm{ADD}}$ be the *concomitant* of $Z_{i:N}^{\mathrm{ADD}}$; i.e., δ_i^{ADD} are sorted according to $Z_{i:N}^{\mathrm{ADD}}$, meaning that $\delta_{[i:N]}^{\mathrm{ADD}} = \delta_j$ if $Z_{i:N}^{\mathrm{ADD}} = Z_j^{\mathrm{ADD}}$. The *survival function of detection points* is given by

$$1 - \hat{F}_N^{\mathrm{ADD}}(t) := \hat{S}^{\mathrm{ADD}}(t) = \prod_{i=1}^{N} (1 - \frac{\delta_{[i:N]}^{\mathrm{ADD}}}{N - i + 1})^{\mathbb{1}(Z_{i:N}^{\mathrm{ADD}} \leq t)}, \tag{47}$$

where $t \in \mathbb{R}_{\geq 0}$ This is equivalent to the convention in the main text, which can be confirmed by canceling factors telescopically. In case of ties, treat a death as if it occurs before a censoring at the same time point. Our KM-ARL is formally

defined as

$$\hat{M}_T^{(\mathrm{KM})} = \int_0^b t \, d\hat{F}_N^{\mathrm{ADD}}, \tag{48}$$

where $a \in \mathbb{R}_{\geq 0}$ is arbitrary.

Theorem B.3 (Finite-sample bias bounds for KM-ADD (Thm. 4.2))**.** *We idealize the time index as continuous without loss of generality, for technical convenience. Consider only sequences with $\nu_i < \infty$ and $\Delta\tau_i \geq 0$ in the dataset. Let $F^{\mathrm{ADD}}, G^{\mathrm{ADD}}$, and H^{ADD} be the CDFs of $\Delta\tau$, C^{ADD}, and Z^{ADD}, respectively. Assume that (i) F^{ADD} and G^{ADD} do not have common discontinuities, (ii) $\Delta\tau$ and C^{ADD} are independent, known as the independent censoring (or non-informative censoring), and (iii) H^{ADD} is continuous. Then, we have for any $b \in \mathbb{R}_{\geq 0}$*

$$-\int_0^b t \, G^{\mathrm{ADD}}(t) \, H^{\mathrm{ADD}}(t)^{N-1} \, F^{\mathrm{ADD}}(dt) \tag{49}$$

$$\leq \mathcal{B}_{\mathrm{FS}}(\hat{M}_T^{(\mathrm{KM})}) \leq \tag{50}$$

$$\int_0^b b \, G^{\mathrm{ADD}}(t) \, H^{\mathrm{ADD}}(t)^{N-1} \, F^{\mathrm{ADD}}(dt), \tag{51}$$

Assumption (i) is a technical requirement and is valid unless pathological situations are considered. In most cases, it holds with probability 1 because we consider continuous time. Note that F^{ADD} and G^{ADD} may have separate discontinuities. Assumption (ii) is the independent censoring assumption discussed above. Assumption (iii) is also a technical requirement.

Proof. Thanks to the setup detailed in this section, the proof is identical to that of Thm. 4.1 with relevant replacements, such as $(\cdot)^{\mathrm{ARL}}$ with $(\cdot)^{\mathrm{ADD}}$. □

B.3. Proof of Thm. 4.3

Theorem B.4 (Truncation bias bounds for KM-ARL (Thm. 4.3))**.** *Define the truncation bias of the LB-ARL as $\mathcal{B}_{\mathrm{TR}}(\hat{\mu}_T^{(\mathrm{LB})}) := \mu_T^{(\mathrm{LB})} - \mu_\infty$, where $\mu_T^{(\mathrm{LB})} := \mathbb{E}[\tau \mid \nu = \infty, \tau \leq T]$ is the true ARL under random sequence lengths. For $a = T_{\max}^*$, we have $\mathcal{B}_{\mathrm{TR}}(\hat{\mu}_T^{(\mathrm{LB})}) \leq \mathcal{B}_{\mathrm{TR}}(\hat{\mu}_T^{(\mathrm{KM})}) \leq 0$.*

Proof. Recall that

$$\mathcal{B}_{\mathrm{TR}}(\hat{\mu}_T^{(\mathrm{KM})}) = \mu_T^{(\mathrm{KM})} - \mu_\infty$$

$$= \int_0^{T_{\max}^*} S^{\mathrm{ARL}}(t)dt - \mathbb{E}[\tau \mid \nu = \infty] \tag{52}$$

$$\mathcal{B}_{\mathrm{TR}}(\hat{\mu}_T^{(\mathrm{LB})}) = \mu_T^{(\mathrm{LB})} - \mu_\infty$$

$$= \mathbb{E}[\tau \mid \nu = \infty, \tau \leq T] - \mathbb{E}[\tau \mid \nu = \infty]. \tag{53}$$

Since

$$\int_0^\infty S^{\mathrm{ARL}}(t)dt \tag{54}$$

$$= \int_0^\infty P(\tau > t \mid \nu = \infty)dt \tag{55}$$

$$= \int_0^\infty \mathbb{E}[\mathbb{1}(\tau > t) \mid \nu = \infty]dt \tag{56}$$

$$= \mathbb{E}[\int_0^\infty \mathbb{1}(\tau > t)dt \mid \nu = \infty] \tag{57}$$

$$= \mathbb{E}[\tau \mid \nu = \infty], \tag{58}$$

we have $\mathcal{B}_{\mathrm{TR}}(\hat{\mu}_T^{(\mathrm{KM})}) = -\int_{T^*_{\max}}^{\infty} S^{\mathrm{ARL}}(t) \leq 0$. Below, we show that $\mathcal{B}_{\mathrm{TR}}(\hat{\mu}_T^{(\mathrm{LB})}) \leq \mathcal{B}_{\mathrm{TR}}(\hat{\mu}_T^{(\mathrm{KM})})$. Since

$$
\begin{aligned}
&\mu_T^{(\mathrm{KM})} \\
&= \int_0^{T^*_{\max}} S^{\mathrm{ARL}}(t)dt && (59)\\
&= \int_0^{T^*_{\max}} P(\tau > t \mid \nu = \infty)dt && (60)\\
&= \int_0^{T^*_{\max}} \mathbb{E}[\mathbb{1}(\tau > t) \mid \nu = \infty]dt && (61)\\
&= \mathbb{E}[\int_0^{T^*_{\max}} \mathbb{1}(\tau > t)dt \mid \nu = \infty] && (62)\\
&= P(\tau \leq T^*_{\max} \mid \nu = \infty) \\
&\quad \times \mathbb{E}[\int_0^{T^*_{\max}} \mathbb{1}(\tau > t)dt \mid \nu = \infty, \tau \leq T^*_{\max}] \\
&\quad + P(\tau > T^*_{\max} \mid \nu = \infty) \\
&\quad \times \mathbb{E}[\int_0^{T^*_{\max}} \mathbb{1}(\tau > t)dt \mid \nu = \infty, \tau > T^*_{\max}] && (63)\\
&= P(\tau \leq T^*_{\max} \mid \nu = \infty)\mathbb{E}[\tau \mid \nu = \infty, \tau \leq T^*_{\max}] \\
&\quad + P(\tau > T^*_{\max} \mid \nu = \infty)\mathbb{E}[T^*_{\max} \mid \nu = \infty, \tau > T^*_{\max}] && (64)\\
&= P(\tau \leq T^*_{\max} \mid \nu = \infty)\mathbb{E}[\tau \mid \nu = \infty, \tau \leq T^*_{\max}] \\
&\quad + P(\tau > T^*_{\max} \mid \nu = \infty)T^*_{\max} && (65)\\
&= (1 - P(\tau > T^*_{\max} \mid \nu = \infty))\mathbb{E}[\tau \mid \nu = \infty, \tau \leq T^*_{\max}] \\
&\quad + P(\tau > T^*_{\max} \mid \nu = \infty)T^*_{\max} && (66)\\
&= \mathbb{E}[\tau \mid \nu = \infty, \tau \leq T^*_{\max}] \\
&\quad + P(\tau > T^*_{\max} \mid \nu = \infty) \\
&\quad \times (T^*_{\max} - \mathbb{E}[\tau \mid \nu = \infty, \tau \leq T^*_{\max}]) && (67)
\end{aligned}
$$

we have

$$
\begin{aligned}
&\mathcal{B}_{\mathrm{TR}}(\hat{\mu}_T^{(\mathrm{KM})}) - \mathcal{B}_{\mathrm{TR}}(\hat{\mu}_T^{(\mathrm{LB})}) \\
&= \mu_T^{(\mathrm{KM})} - \mu_T^{(\mathrm{LB})} && (68)\\
&= \mathbb{E}[\tau \mid \nu = \infty, \tau \leq T^*_{\max}] \\
&\quad + P(\tau > T^*_{\max} \mid \nu = \infty) \\
&\quad \times (T^*_{\max} - \mathbb{E}[\tau \mid \nu = \infty, \tau \leq T^*_{\max}]) \\
&\quad - \mathbb{E}[\tau \mid \nu = \infty, \tau \leq T] && (69)\\
&= \mathbb{E}[\tau \mid \nu = \infty, \tau \leq T^*_{\max}] - \mathbb{E}[\tau \mid \nu = \infty, \tau \leq T] \\
&\quad + P(\tau > T^*_{\max} \mid \nu = \infty) \\
&\quad \times (T^*_{\max} - \mathbb{E}[\tau \mid \nu = \infty, \tau \leq T^*_{\max}]). && (70)
\end{aligned}
$$

Finally, we can see that $\mathbb{E}[\tau \mid \nu = \infty, \tau \leq T^*_{\max}] - \mathbb{E}[\tau \mid \nu = \infty, \tau \leq T] \geq 0$ (Lem. B.5), and $T^*_{\max} - \mathbb{E}[\tau \mid \nu = \infty, \tau \leq T^*_{\max}] \geq 0$; hence, $\mathcal{B}_{\mathrm{TR}}(\hat{\mu}_T^{(\mathrm{KM})}) - \mathcal{B}_{\mathrm{TR}}(\hat{\mu}_T^{(\mathrm{LB})}) \geq 0$. □

Lemma B.5. $\mathbb{E}[\tau \mid \nu = \infty, \tau \leq T^*_{\max}] - \mathbb{E}[\tau \mid \nu = \infty, \tau \leq T] \geq 0$.

Proof. Without loss of generality, this inequality is equivalent to $\mathbb{E}[\tau \mid 0 \leq \tau \leq T^*_{\max}] - \mathbb{E}[\tau \mid 0 \leq \tau \leq T] \geq 0$, where $T \leq T^*_{\max} (\in \mathbb{R})$ holds a.s., and τ and T are independent. For notational simplicity, we will show that

$$\mathbb{E}_{X,Y}[X|X \leq y_{\max}] - \mathbb{E}_{X,Y}[X|X \leq Y] \geq 0, \tag{71}$$

where X is a non-negative random variable, Y is a non-negative random variable with $0 \leq Y \leq y_{\max}$ a.s., $y_{\max}(\geq 0)$ is a real constant, and X and Y are independent.

Write F_X for the CDF of X, and let

$$g(x) := P(Y \geq x), \quad 0 \leq x \leq y_{\max}. \tag{72}$$

Because $Y \in [0, y_{\max}]$, g is non-increasing in x, while $g(x) = 0$ for $x > y_{\max}$. Thus, we have

$$P(X \leq Y) = \mathbb{E}\left[\mathbb{1}(X \leq Y)\right] = \int_0^\infty P(Y \geq x) dF_X(x) = \int_0^{y_{\max}} g(x) dF_X(x), \tag{73}$$

where we used the independence of X and Y. Additionally, we have

$$\mathbb{E}\left[X\mathbb{1}(X \leq Y)\right] = \int_0^\infty x P(Y \geq x) dF_X(x) = \int_0^{y_{\max}} x g(x) dF_X(x). \tag{74}$$

Therefore,

$$\mathbb{E}[X \mid X \leq Y] = \frac{\mathbb{E}\left[X\mathbb{1}(X \leq Y)\right]}{P(X \leq Y)} = \frac{\int_0^{y_{\max}} x g(x) dF_X(x)}{\int_0^{y_{\max}} g(x) dF_X(x)}. \tag{75}$$

On the other hand, we can similarly derive

$$\mathbb{E}\left[X \mid X \leq y_{\max}\right] = \frac{\int_0^{y_{\max}} x dF_X(x)}{\int_0^{y_{\max}} dF_X(x)}. \tag{76}$$

Next, let ν be the finite measure on $[0, y_{\max}]$ given by $d\nu(x) = dF_X(x)$, and define another measure μ by

$$d\mu(x) = g(x) d\nu(x), \quad 0 \leq x \leq y_{\max}. \tag{77}$$

Note that $g(x)$ is non-increasing in x and that x is increasing in x. Thus, Eqs. (75) and (76) can be rewritten as

$$\mathbb{E}[X \mid X \leq Y] = \frac{\int x d\mu(x)}{\int d\mu(x)}, \quad \mathbb{E}\left[X \mid X \leq y_{\max}\right] = \frac{\int x d\nu(x)}{\int d\nu(x)}. \tag{78}$$

Consider the covariance under the normalized measure proportional to ν :

$$\mathrm{Cov}_\nu(x, g(x)) = \mathbb{E}_\nu[X g(X)] - \mathbb{E}_\nu[X]\, \mathbb{E}_\nu[g(X)] \leq 0, \tag{79}$$

because x is increasing and $g(x)$ is decreasing (Lem. B.6). Therefore,

$$\left(\frac{\int x g(x) d\nu(x)}{\int d\nu(x)}\right) \leq \left(\frac{\int x d\nu(x)}{\int d\nu(x)}\right)\left(\frac{\int g(x) d\nu(x)}{\int d\nu(x)}\right). \tag{80}$$

This is equivalent to

$$\frac{\int x g(x) d\nu(x)}{\int g(x) d\nu(x)} \leq \frac{\int x d\nu(x)}{\int d\nu(x)}. \tag{81}$$

In other words,

$$\mathbb{E}[X \mid X \leq Y] \leq \mathbb{E}\left[X \mid X \leq y_{\max}\right]. \tag{82}$$

This concludes the proof. □

Lemma B.6. *Let X be a real random variable, f an increasing functions, and h and non-increasing function. Then, it holds that* $\mathrm{Cov}(f(X), h(X)) \leq 0$.

Proof. Take two i.i.d. copies, X and X', with the same distribution. Let us consider the random quantity

$$\left(f(X) - f\left(X'\right)\right)\left(h(X) - h\left(X'\right)\right). \tag{83}$$

Expanding its expectation, we have

$$\begin{aligned}\mathbb{E}\left[\left(f(X)-f\left(X'\right)\right)\left(h(X)-h\left(X'\right)\right)\right] =&\mathbb{E}[f(X)h(X)]+\mathbb{E}\left[f\left(X'\right)h\left(X'\right)\right]\\ &-\mathbb{E}\left[f(X)h\left(X'\right)\right]-\mathbb{E}\left[f\left(X'\right)h(X)\right] \quad (84)\\ =&2\mathbb{E}[f(X)h(X)]-2\mathbb{E}[f(X)]\mathbb{E}[h(X)] \quad (85)\\ =&2\mathrm{Cov}(f(X),h(X)). \quad (86)\end{aligned}$$

Thus, we have the identity

$$\mathrm{Cov}(f(X),h(X))=\frac{1}{2}\mathbb{E}\left[\left(f(X)-f\left(X'\right)\right)\left(h(X)-h\left(X'\right)\right)\right]. \quad (87)$$

Now use the fact that: f is increasing, and h is non-increasing. Fix any x, y in the support. Then:

- If $x > y$, then $f(x) > f(y)$ and $h(x) \leq h(y)$, so $(f(x) - f(y)) > 0$ and $(h(x) - h(y)) \leq 0$, hence $(f(x) - f(y))(h(x) - h(y)) \leq 0$.
- If $x < y$, then $f(x) < f(y)$ and $h(x) \geq h(y)$, so $(f(x) - f(y)) < 0$ and $(h(x) - h(y)) \geq 0$, again $(f(x) - f(y))(h(x) - h(y)) \leq 0$.
- If $x = y$, the product is 0.

Thus, we have

$$(f(x)-f(y))(h(x)-h(y))\leq 0 \text{ for all } x,y. \quad (88)$$

Therefore, when X and X' are i.i.d.,

$$\left(f(X)-f\left(X'\right)\right)\left(h(X)-h\left(X'\right)\right)\leq 0 \text{ a.s.} \quad (89)$$

Taking expectations,

$$\mathbb{E}\left[\left(f(X)-f\left(X'\right)\right)\left(h(X)-h\left(X'\right)\right)\right]\leq 0. \quad (90)$$

Plug this back into Eq. (87), we have

$$\mathrm{Cov}(f(X),h(X))=\frac{1}{2}\mathbb{E}\left[\left(f(X)-f\left(X'\right)\right)\left(h(X)-h\left(X'\right)\right)\right]\leq 0. \quad (91)$$

□

B.4. Proof of Theorem 4.4

Theorem B.7 (Truncation bias bound for KM-ADD)**.** *Define the truncation bias of the LB-ADD as* $\mathcal{B}_{\mathrm{TR}}(\hat{M}_T^{(\mathrm{LB})}) := M_T^{(\mathrm{LB})} - M_\infty$*, where* $M_T^{(\mathrm{LB})} := \mathbb{E}[\Delta\tau \mid \nu < \infty, 0 \leq \Delta\tau \leq \Delta T]$ *is the true ADD under random sequence lengths. For* $b = \Delta T^*_{\max}$*, we have* $\mathcal{B}_{\mathrm{TR}}(\hat{M}_T^{(\mathrm{LB})}) \leq \mathcal{B}_{\mathrm{TR}}(\hat{M}_T^{(\mathrm{KM})}) \leq 0$*, under the independent censoring assumption in Thm. 4.2.*

Proof. Recall that

$$\begin{aligned}&\mathcal{B}_{\mathrm{TR}}(\hat{M}_T^{(\mathrm{KM})}) = M_T^{(\mathrm{KM})} - M_\infty\\ =&\int_0^{\Delta T^*_{\max}} S^{\mathrm{ADD}}(t)dt - \mathbb{E}[\Delta\tau \mid \nu<\infty, \Delta\tau\geq 0] \quad (92)\\ &\mathcal{B}_{\mathrm{TR}}(\hat{M}_T^{(\mathrm{LB})}) = M_T^{(\mathrm{LB})} - M_\infty\\ =&\mathbb{E}[\Delta\tau \mid \nu<\infty, 0\leq\Delta\tau\leq\Delta T] - \mathbb{E}[\Delta\tau \mid \nu<\infty, \Delta\tau\geq 0]. \quad (93)\end{aligned}$$

First, since

$$\int_0^\infty S^{\mathrm{ADD}}(t)dt \tag{94}$$
$$= \int_0^\infty P(\Delta\tau > t \mid \nu < \infty, \Delta\tau \geq 0)dt \tag{95}$$
$$= \int_0^\infty \mathbb{E}[\mathbb{1}(\Delta\tau > t) \mid \nu < \infty, \Delta\tau \geq 0]dt \tag{96}$$
$$= \mathbb{E}[\int_0^\infty \mathbb{1}(\Delta\tau > t)dt \mid \nu < \infty, \Delta\tau \geq 0] \tag{97}$$
$$= \mathbb{E}[\Delta\tau \mid \nu < \infty, \Delta\tau \geq 0], \tag{98}$$

we have $\mathcal{B}_{\mathrm{TR}}(\hat{M}_T^{(\mathrm{KM})}) = -\int_{\Delta T^*_{\max}}^\infty S^{\mathrm{ADD}}(t) \leq 0$.

Second, we will show that $\mathcal{B}_{\mathrm{TR}}(\hat{M}_T^{(\mathrm{LB})}) \leq \mathcal{B}_{\mathrm{TR}}(\hat{M}_T^{(\mathrm{KM})})$. Since

$$M_T^{(\mathrm{KM})}$$
$$= \int_0^{\Delta T^*_{\max}} S^{\mathrm{ADD}}(t)dt \tag{99}$$
$$= \int_0^{\Delta T^*_{\max}} P(\Delta\tau > t \mid \nu < \infty, \Delta\tau \geq 0)dt \tag{100}$$
$$= \int_0^{\Delta T^*_{\max}} \mathbb{E}[\mathbb{1}(\Delta\tau > t) \mid \nu < \infty, \Delta\tau \geq 0]dt \tag{101}$$
$$= \mathbb{E}[\int_0^{\Delta T^*_{\max}} \mathbb{1}(\Delta\tau > t)dt \mid \nu < \infty, \Delta\tau \geq 0] \tag{102}$$
$$\begin{aligned} =& P(\Delta\tau \leq \Delta T^*_{\max} \mid \nu < \infty, \Delta\tau \geq 0) \\ &\times \mathbb{E}[\int_0^{\Delta T^*_{\max}} \mathbb{1}(\Delta\tau > t)dt \mid \nu < \infty, 0 \leq \Delta\tau \leq \Delta T^*_{\max}] \\ &+ P(\Delta\tau > \Delta T^*_{\max} \mid \nu < \infty, \Delta\tau \geq 0) \\ &\times \mathbb{E}[\int_0^{\Delta T^*_{\max}} \mathbb{1}(\Delta\tau > t)dt \mid \nu < \infty, 0 \leq \Delta\tau > \Delta T^*_{\max}] \end{aligned} \tag{103}$$
$$\begin{aligned} =& P(\Delta\tau \leq \Delta T^*_{\max} \mid \nu < \infty, \Delta\tau \geq 0) \\ &\times \mathbb{E}[\Delta\tau \mid \nu < \infty, 0 \leq \Delta\tau \leq \Delta T^*_{\max}] \\ &+ P(\Delta\tau > \Delta T^*_{\max} \mid \nu < \infty, \Delta\tau \geq 0) \\ &\times \mathbb{E}[\Delta T^*_{\max} \mid \nu < \infty, 0 \leq \Delta\tau > \Delta T^*_{\max}] \end{aligned} \tag{104}$$

$$\begin{aligned} =& P(\Delta\tau \leq \Delta T^*_{\max} \mid \nu < \infty, \Delta\tau \geq 0) \\ &\times \mathbb{E}[\Delta\tau \mid \nu < \infty, 0 \leq \Delta\tau \leq \Delta T^*_{\max}] \\ &+ P(\Delta\tau > \Delta T^*_{\max} \mid \nu < \infty)\Delta T^*_{\max} \end{aligned} \tag{105}$$
$$\begin{aligned} =& (1 - P(\Delta\tau > \Delta T^*_{\max} \mid \nu < \infty, \Delta\tau \geq 0)) \\ &\times \mathbb{E}[\Delta\tau \mid \nu < \infty, 0 \leq \Delta\tau \leq \Delta T^*_{\max}] \\ &+ P(\Delta\tau > \Delta T^*_{\max} \mid \nu < \infty, \Delta\tau \geq 0)\Delta T^*_{\max} \end{aligned} \tag{106}$$
$$\begin{aligned} =& \mathbb{E}[\Delta\tau \mid \nu < \infty, 0 \leq \Delta\tau \leq \Delta T^*_{\max}] \\ &+ P(\Delta\tau > \Delta T^*_{\max} \mid \nu < \infty, \Delta\tau \geq 0) \\ &\times (\Delta T^*_{\max} - \mathbb{E}[\Delta\tau \mid \nu < \infty, 0 \leq \Delta\tau \leq \Delta T^*_{\max}]) \end{aligned} \tag{107}$$

we have

$$
\begin{aligned}
&\mathcal{B}_{\mathrm{TR}}(\hat{M}_T^{(\mathrm{KM})}) - \mathcal{B}_{\mathrm{TR}}(\hat{M}_T^{(\mathrm{LB})}) \\
=&M_T^{(\mathrm{KM})} - M_T^{(\mathrm{LB})} && (108)\\
=&\mathbb{E}[\Delta\tau \mid \nu < \infty, 0 \le \Delta\tau \le \Delta T^*_{\max}] \\
&+ P(\Delta\tau > \Delta T^*_{\max} \mid \nu < \infty, \Delta\tau \ge 0) \\
&\times (\Delta T^*_{\max} - \mathbb{E}[\Delta\tau \mid \nu < \infty, 0 \le \Delta\tau \le \Delta T^*_{\max}]) \\
&- \mathbb{E}[\Delta\tau \mid \nu < \infty, 0 \le \Delta\tau \le \Delta T] && (109)\\
=&\mathbb{E}[\Delta\tau \mid \nu < \infty, 0 \le \Delta\tau \le \Delta T^*_{\max}] \\
&- \mathbb{E}[\Delta\tau \mid \nu < \infty, 0 \le \Delta\tau \le \Delta T] \\
&+ P(\Delta\tau > \Delta T^*_{\max} \mid \nu < \infty, \Delta\tau \ge 0) \\
&\times (\Delta T^*_{\max} - \mathbb{E}[\Delta\tau \mid \nu < \infty, 0 \le \Delta\tau \le \Delta T^*_{\max}]). && (110)
\end{aligned}
$$

Finally, we can see that $\mathbb{E}[\Delta\tau \mid \nu < \infty, 0 \le \Delta\tau \le \Delta T^*_{\max}] - \mathbb{E}[\Delta\tau \mid \nu < \infty, 0 \le \Delta\tau \le \Delta T] \ge 0$ (Lem. B.8), and $\Delta T^*_{\max} - \mathbb{E}[\Delta\tau \mid \nu < \infty, 0 \le \Delta\tau \le \Delta T^*_{\max}] \ge 0$; hence, $\mathcal{B}_{\mathrm{TR}}(\hat{M}_T^{(\mathrm{KM})}) - \mathcal{B}_{\mathrm{TR}}(\hat{M}_T^{(\mathrm{LB})}) \ge 0$.

□

Lemma B.8. *Under the independent censoring assumption, it holds that* $\mathbb{E}[\Delta\tau \mid \nu < \infty, 0 \le \Delta\tau \le \Delta T^*_{\max}] - \mathbb{E}[\Delta\tau \mid \nu < \infty, 0 \le \Delta\tau \le \Delta T] \ge 0$.

Proof. Without loss of generality, this inequality is equivalent to $\mathbb{E}[\Delta\tau \mid 0 \le \Delta\tau \le \Delta T^*_{\max}] - \mathbb{E}[\Delta\tau \mid 0 \le \Delta\tau \le \Delta T] \ge 0$, where $\Delta T \le \Delta T^*_{\max} (\in \mathbb{R})$ holds a.s. $\Delta\tau$ and $\Delta T (= C^{\mathrm{ADD}})$ are independent because of the independent censoring assumption. This inequality can be rewritten as

$$
\mathbb{E}_{X,Y}[X|X \le y_{\max}] - \mathbb{E}_{X,Y}[X|X \le Y] \ge 0, \tag{111}
$$

where X is a non-negative random variable, Y is a non-negative random variable with $0 \le Y \le y_{\max}$ a.s., $y_{\max} (\ge 0)$ is a real constant, and X and Y are independent. Because Ineq. (111) coincides with Ineq. (71), the proof follows the same steps as Lem. B.5. □

C. Supplementary Experimental Results

C.1. Example of Survival Curve of Detection Points

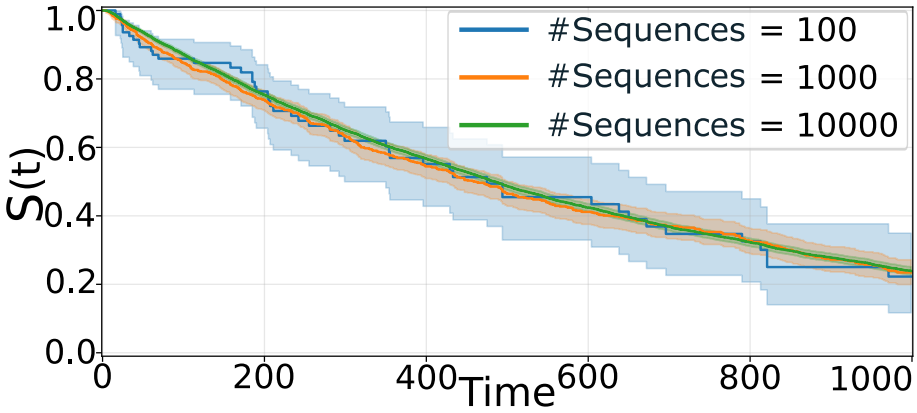


Figure 6. **Estimated survival functions of detection points.** The shaded area represents the standard error of the mean.

Survival functions of detection points are shown in Fig. 6. The experiment uses the Gaussian process dataset at three dataset sizes. The QCD model used is the GSR procedure with ground-truth statistics, evaluated under a given threshold. Changepoint locations are sampled from a geometric distribution with the success probability of $p = 0.001$. Error bars represent the standard error of the mean.

C.2. Complete Ablation of Fig. 4

Fig. 7 presents the complete ablation study corresponding to Fig. 4, demonstrating that KM-ARL and KM-ADD yield more accurate estimates of the true ARL-ADD curve, even under limited and irregular sequence lengths. The gray areas indicate regions of extrapolation (excluding the true ARL), where ARLs cannot be estimated due to the absence of data, unless additional assumptions are imposed on the underlying distribution.

C.3. Gaussian Process Dataset

We provide additional experimental results on the Gaussian process dataset. Fig. 8 is an example sequence. Fig. 9 shows temporal statistics of a sequence in our Gaussian process dataset. Fig. 10 presents the ARL-ADD tradeoff curves of the QCD models evaluated on a Gaussian process dataset, while we use the WISDM Actitracker dataset in Fig. 1 in the main text. It demonstrates that our KM-ARL and KM-ADD are more robust to irregular lengths than the conventional LB-ARL and LB-ADD. Note that some curves in Fig. 10(a) are non-monotonic because the sequences used to compute the LB-ARL and LB-ADD differ drastically across different thresholds, causing unstable estimates. For example, short sequences are excluded from the computation of the LB metrics when the threshold is high. This issue does not arise for our KME-based metrics, contributing to their robustness to finite and irregular sequence lengths.

C.4. Poisson Process Dataset

We provide experimental results on the Poisson process dataset. The pre-change and post-change Poisson processes have the mean of 1 and 4, respectively. We use two types of changepoint distributions: geometric and uniform. Fig. 8 is an example sequence from our Poisson process dataset. Fig. 12 shows statistics of a sequence in our Poisson process dataset. Fig. 13 presents the ARL-ADD curve evaluated on the Poisson process dataset, demonstrating that the KM-ARL and KM-ADD provide more accurate estimates of the true ARL-ADD curve, even when sequence lengths are limited and irregular.

C.5. WISDM Actitracker Dataset

We provide additional experimental results on the WISDM Actitracker dataset. Fig. 14 is the original figure of Fig. 1.

While we use the machine-labeled subset in the main text, Fig. 15 presents the ARL-ADD curves of QCD models evaluated on the user-labeled subset ("labeled" subset) of the WISDM Actitracker dataset, which is about $10^3 (= 51326/83)$ times

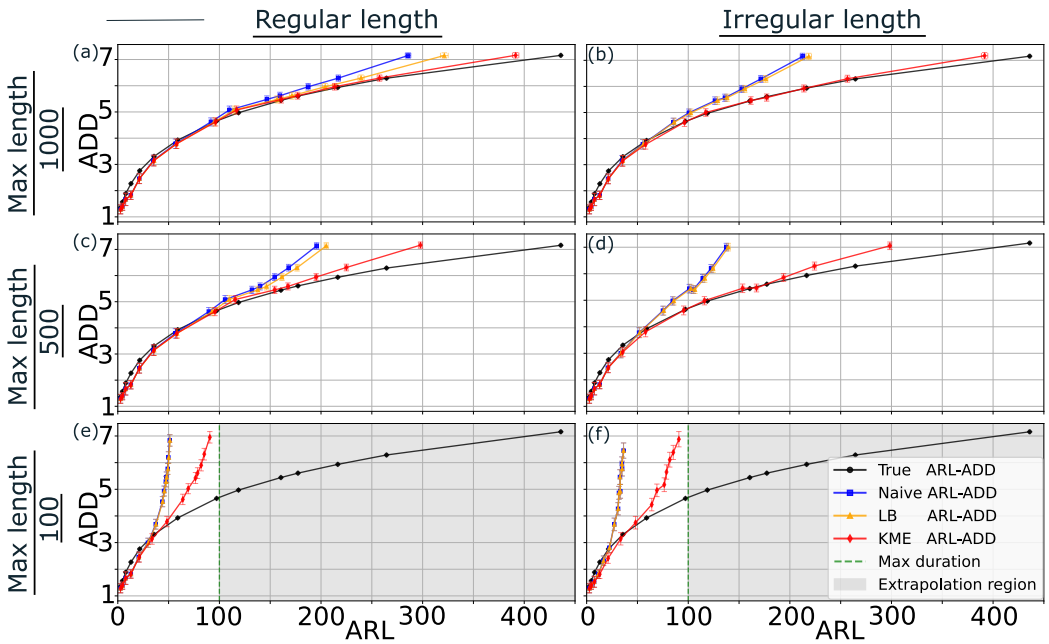


Figure 7. **Complete ablation study of Fig. 4.** The KM-ARL and KM-ADD provide more accurate estimates of the true ARL-ADD curve, even when sequence lengths are limited and irregular. The experiments use the Gaussian process dataset, comprising 10000 sequences. The employed QCD algorithm is the GSR procedure using ground-truth statistics. Changepoint locations are sampled uniformly. 50% of sequences contain a changepoint. Error bars represent the standard error of the mean. The gray areas indicate regions of extrapolation (excluding the true ARL), where ARLs cannot be estimated due to the absence of data, unless additional assumptions are imposed on the underlying distribution. **(a) Sequence length is** 1000**. (b) Sequence lengths vary irregularly in the range** $[100, 1000]$**. (c) Sequence length is** 500**. (d) Sequence lengths vary irregularly in the range** $[50, 500]$**. (e) Sequence length is** 100**. (f) Sequence lengths vary irregularly in the range** $[10, 100]$**.**

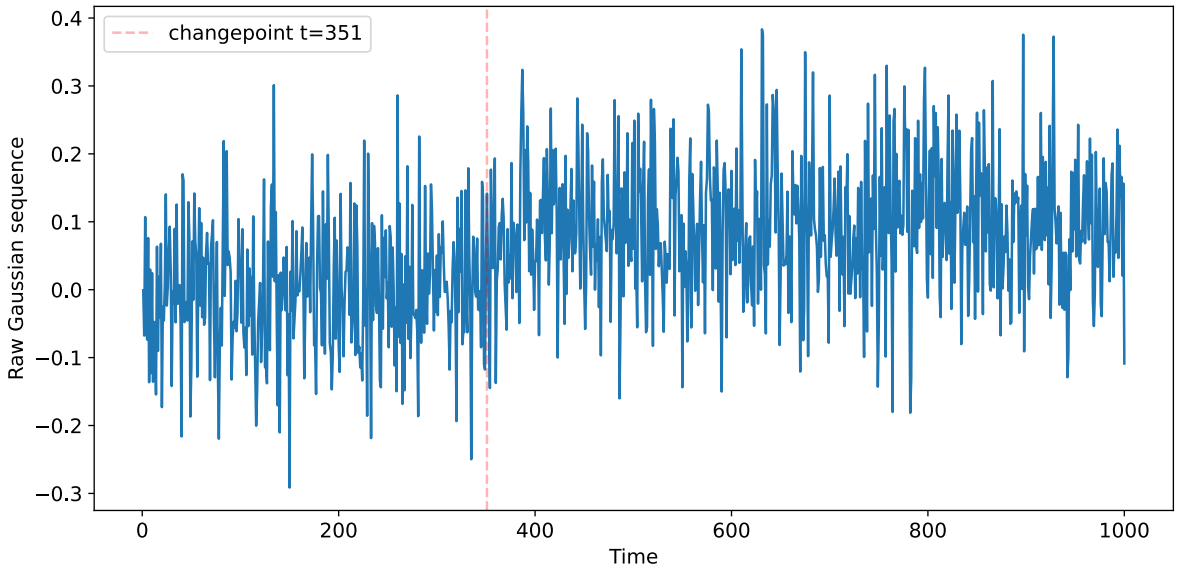


Figure 8. **Example sequence in Gaussian process dataset.**

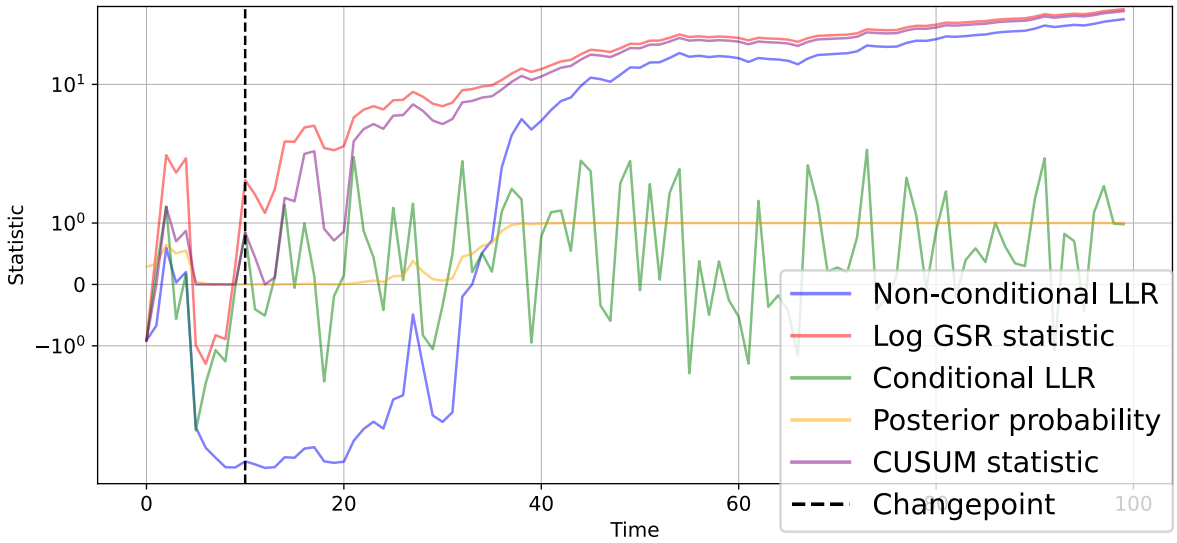


Figure 9. **Example statistics in Gaussian process dataset.** The observation interval is $\Delta t = 1$. LLR is short for log-likelihood ratio $\log(f(X^{(0,t)})/g(X^{(0,t)}))$, where f and g are the pre- and post-change density, respectively. Non-conditional LLR is given by $\log(f(X^{(0,t)})/g(X^{(0,t)}))$, while Conditional LLR is given by $\log(f(X^{(t)} \mid X^{(0,t-1)}))/g(X^{(t)} \mid X^{(0,t-1)}))$. Posterior probability means $g(X^{(0,t)})$.

smaller than the machine-labeled subset ("unlabeled" subset) of the WISDM Actitracker dataset (see Tab. 2 for statistics). Evaluation is challenging on this subset because the number of sequences is limited to only 83. For such small datasets, we recommend using min-max-based metrics, such as the box plot, rather than average-based metrics, such as the ARL and ADD.

C.6. ARL-ADD with CUSUM

While we use the GSR procedure in the main text, we provide additional experimental results obtained with the CUSUM procedure (Page, 1954) with ground-truth statistics. The ARL-ADD curves are provided in Fig. 16 and support our findings in the main text. The CUSUM procedure is evaluated under various thresholds.

C.7. ARL-ADD with Geometric Changepoint Distribution

We provide additional experimental results obtained with the geometric changepoint distribution, instead of the uniform distribution. We evaluate the ARL-ADD tradeoff curve on a Gaussian process dataset. The results are provided in Fig. 17 and support our findings in the main text.

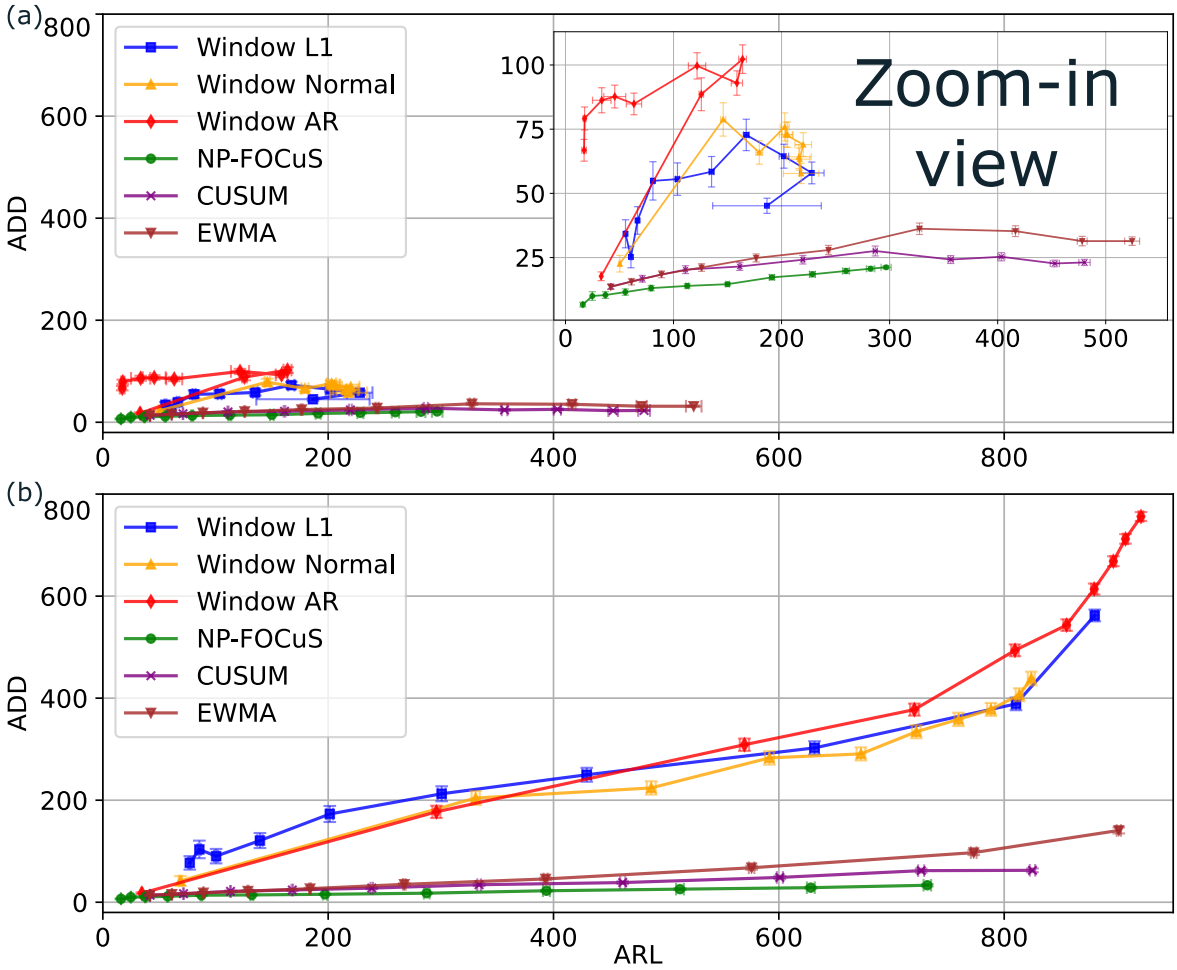


Figure 10. **Evaluation of QCD models on Gaussian process dataset. (a) LB-ARL and LB-ADD. (b) KM-ARL and KM-ADD.** Our KM-ARL and KM-ADD are more robust to irregular lengths than the conventional LB-ARL and LB-ADD. The experiments use the Gaussian process dataset, comprising 10000 sequences. Changepoint locations are sampled uniformly. 50% of the sequences contain a changepoint. Sequence lengths vary irregularly in the range $[100, 1000]$. Note that some curves in (a) are non-monotonic because the sequences used to compute the LB-ARL and LB-ADD differ drastically across different thresholds, causing unstable estimates. For example, short sequences are excluded from the computation of the LB metrics when the threshold is high. This issue does not arise for our KME-based metrics, contributing to their robustness to finite and irregular sequence lengths.

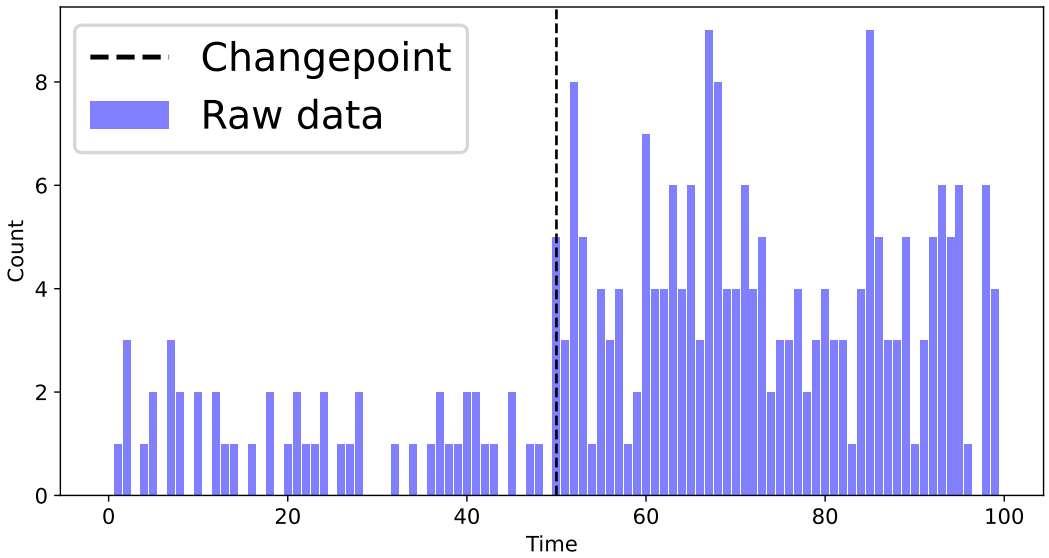


Figure 11. **Example sequence in Poisson process dataset.**

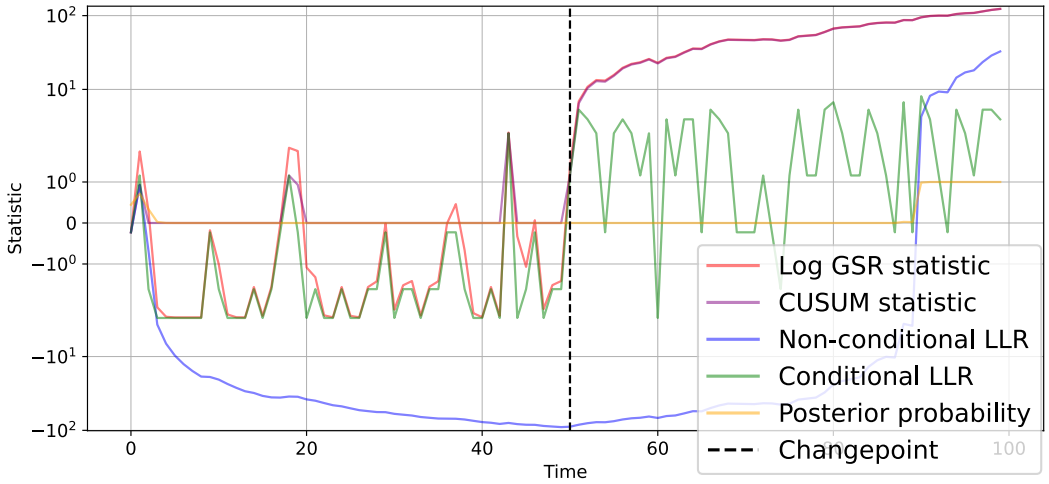


Figure 12. **Example statistics in Poisson process dataset.** Observation interval is $\Delta t = 1$. LLR is short for log-likelihood ratio $\log(f(X^{(0,t)})/g(X^{(0,t)}))$, where f and g are the pre- and post-change density, respectively. Non-conditional LLR is given by $\log(f(X^{(0,t)})/g(X^{(0,t)}))$, while Conditional LLR is given by $\log(f(X^{(t)} \mid X^{(0,t-1)}))/g(X^{(t)} \mid X^{(0,t-1)}))$. Posterior probability means $g(X^{(0,t)})$.

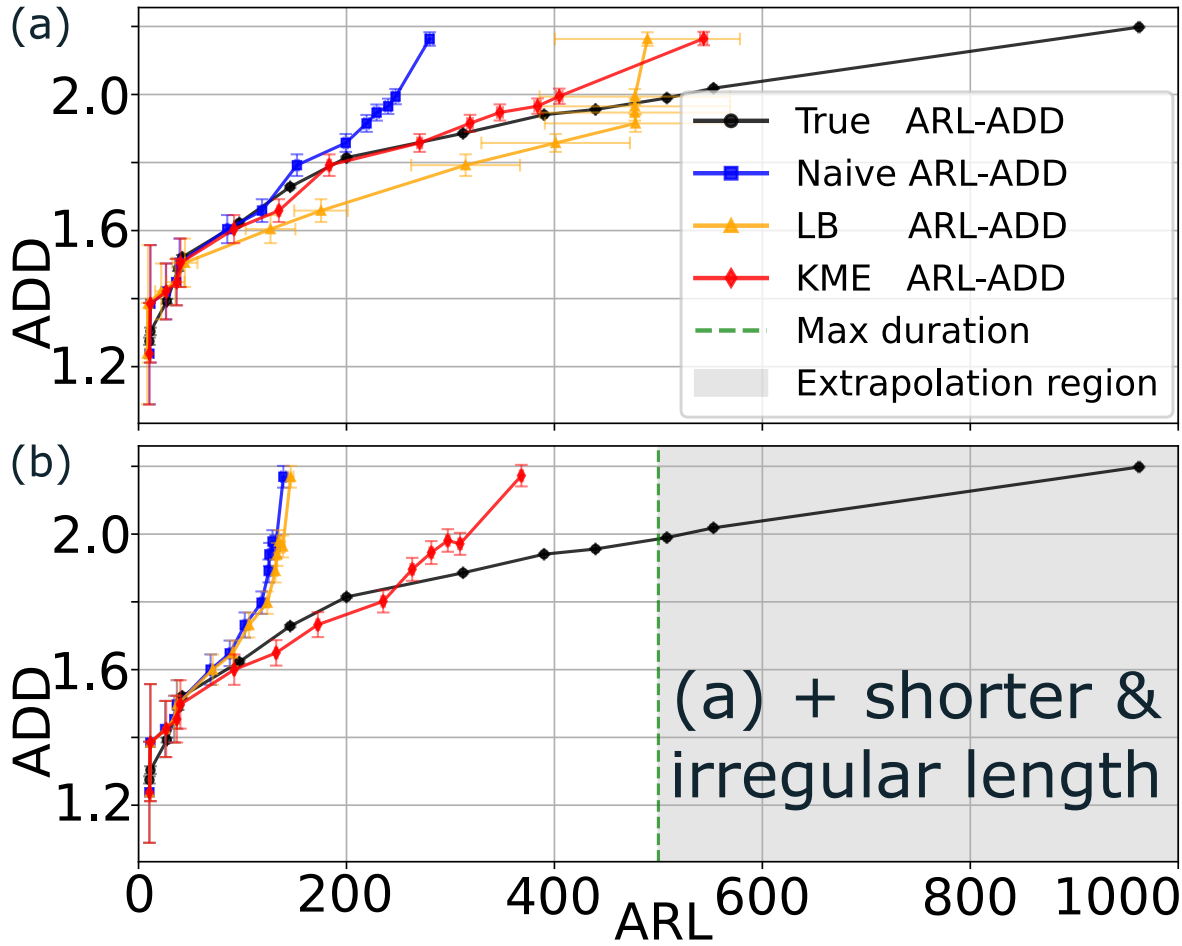


Figure 13. **ARL-ADD tradeoff curves on Poisson process dataset.** The KM-ARL and KM-ADD provide more accurate estimates of the true ARL-ADD curve, even when sequence lengths are limited and irregular. **(a) Sequence length is** 1000**. (b) Sequence lengths vary irregularly in the range** $[50, 500]$**.** The experiments use the Gaussian process dataset, comprising 10000 sequences. The employed QCD algorithm is the GSR procedure using ground-truth statistics. Changepoint locations are sampled uniformly. 50% of sequences contain a changepoint. Error bars represent the standard error of the mean. In (b), $\mathrm{ARL} > 500$ (gray area) indicates a region of extrapolation (excluding the true ARL), where ARLs cannot be estimated due to the absence of data, unless additional assumptions are imposed on the underlying distribution.

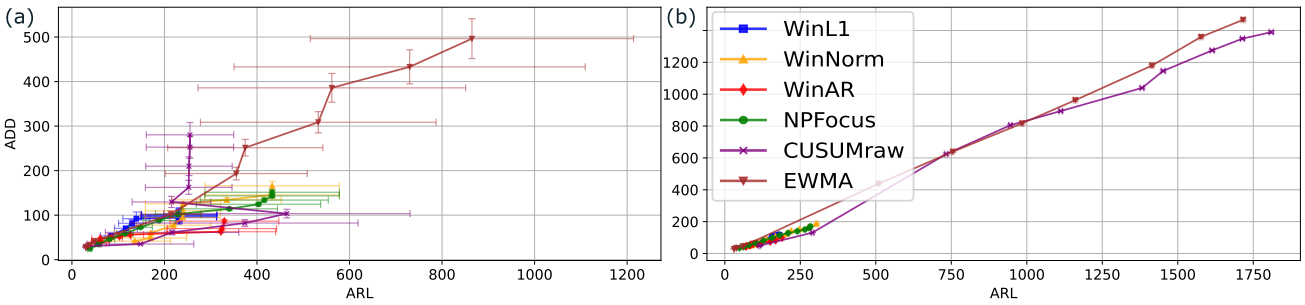


Figure 14. **Original figure of Fig. 1.** Our KM-ARL and KM-ADD are more interpretable than the conventional LB-ARL and LB-ADD.

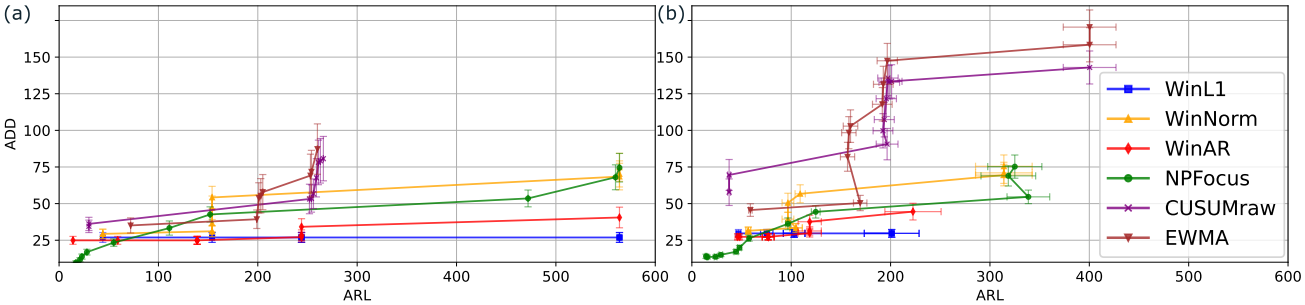


Figure 15. **Evaluation of QCD models on real-world dataset (user-labeled subset of WISDM Actitracker).** The user-labeled subset ("labeled" subset) of the WISDM Actitracker is used, which is about $10^3 (= 51326/83)$ times smaller than the machine-labeled subset ("unlabeled" subset) of the WISDM Actitracker dataset (see Tab. 2 for statistics). Evaluation is challenging on this subset because the number of sequences is limited to only 83. For such small datasets, we recommend using min-max-based metrics, such as the box plot, rather than average-based metrics, such as the ARL and ADD.

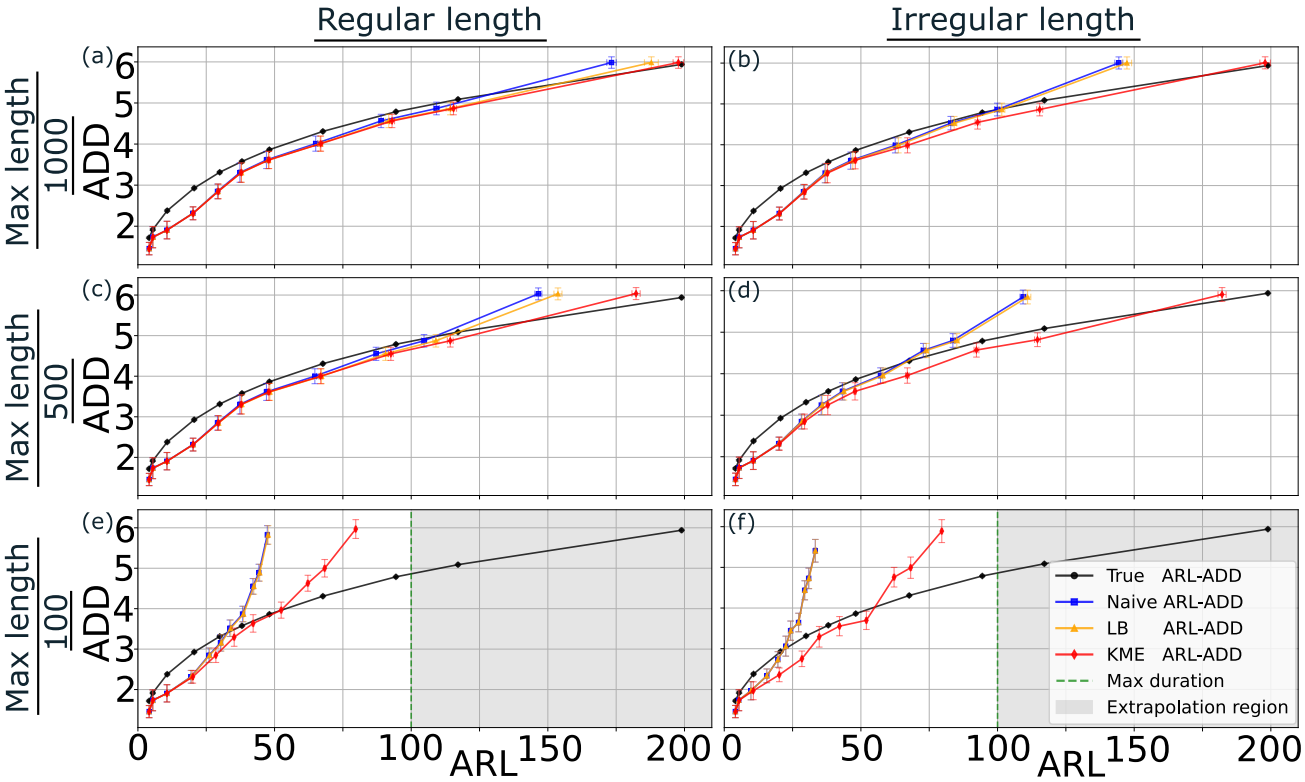


Figure 16. **ARL-ADD curve of CUSUM procedure with ground-truth statistics.** The results closely parallel those obtained with the GSR procedure in the main text: The KM-ARL and KM-ADD provide more accurate estimates of the true ARL-ADD curve, even when sequence lengths are limited and irregular. CUSUM is evaluated under various thresholds. The experiments use the Gaussian process dataset, comprising 10000 sequences. Changepoint locations are sampled uniformly. 50% of sequences contain a changepoint. Error bars represent the standard error of the mean. The gray areas indicate regions of extrapolation (excluding the true ARL), where ARLs cannot be estimated due to the absence of data, unless additional assumptions are imposed on the underlying distribution. **(a) Sequence length is** 1000**. (b) Sequence lengths vary irregularly in the range** $[100, 1000]$**. (c) Sequence length is** 500**. (d) Sequence lengths vary irregularly in the range** $[50, 500]$**. (e) Sequence length is** 100**. (f) Sequence lengths vary irregularly in the range** $[10, 100]$**.**

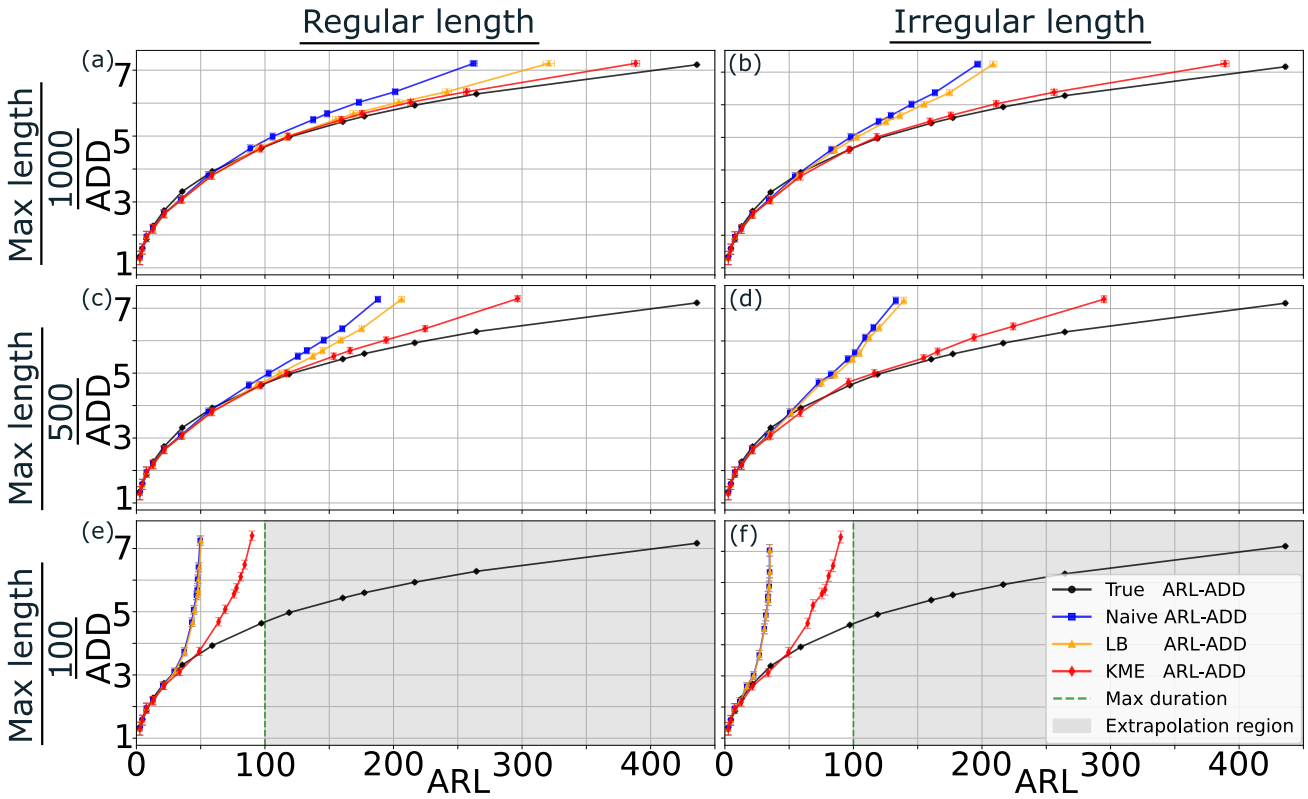


Figure 17. **ARL-ADD curve on Gaussian process dataset with geometric changepoint distribution.** The results closely parallel those obtained with the uniform changepoint distribution in the main text: The KM-ARL and KM-ADD provide more accurate estimates of the true ARL-ADD curve, even when sequence lengths are limited and irregular. The experiments use the Gaussian process dataset, comprising 10000 sequences. The employed QCD algorithm is the GSR procedure using ground-truth statistics. Changepoint locations are sampled from a geometric distribution with the success probability $p = 0.001$. Error bars represent the standard error of the mean. The gray areas indicate regions of extrapolation (excluding the true ARL), where ARLs cannot be estimated due to the absence of data, unless additional assumptions are imposed on the underlying distribution. **(a) Sequence length is** 1000**. (b) Sequence lengths vary irregularly in the range** $[100, 1000]$**. (c) Sequence length is** 500**. (d) Sequence lengths vary irregularly in the range** $[50, 500]$**. (e) Sequence length is** 100**. (f) Sequence lengths vary irregularly in the range** $[10, 100]$**.**

D. Detailed Experimental Settings

Runtime. The computational costs of metric evaluation, including KM-ARL and KM-ADD, are negligible in our experiments compared with the time required to run the QCD algorithms. Computing a single data point in the figures takes from a few seconds to several hours, depending on the sequence length, the number of sequences, and the QCD algorithm used.

Computing Infrastructure. We use Python 3.11.9 (Van Rossum & Drake, 2009), Numpy 1.26.4 (Harris et al., 2020), lifelines 0.30.0 (Davidson-Pilon, 2019), changepoint-online 1.2.1 (Romano et al., 2024b), ruptures 1.1.9 (Truong et al., 2020), ocpdet 0.0.6 (Khamesi, 2022) (we pick up relevant implementations only (`ocpdet/CUSUM.py` and `ocpdet/EWMA.py`), and our implementation does not require TensorFlow (Abadi et al., 2015), which is required when installing ocpdet.). PyTorch 2.6.0 (Paszke et al., 2019) (used only for saving and loading the Gaussian and Poisson datasets and can be replaced with Numpy etc.), The operating system (OS) is Ubuntu 20.04. Intel(R) Core(TM) i9-7980XE CPU @ 2.60 GHz (18 cores) are used. The total random access memory (RAM) size of our server is 125 GBs. We do not use GPUs.

D.1. Online QCD Models

To see a practical relevance of our KME-based metrics, we evaluate six QCD models on the simulated Gaussian process dataset (App. C.3) and real-world WISDM Actitracker dataset (main text and App. C.5). We use the following online QCD models: Window L1 (Bai, 1995), Window Normal (Lavielle, 1999; Lavielle & Teyssiere, 2006), Window AR (Bai, 2000), non-parametric focused changepoint detection (NP-FOCuS) (Romano et al., 2024a), CUSUM (Page, 1954), and exponentially weighted moving average (EWMA) (Roberts, 1959). For Window L1, Window Normal, Window AR, the discrepancy measures for detection are derived from *cost functions* as specified in (Truong et al., 2020).

Window L1 detects changes in the median and uses the L^1 norm for the cost function. Window Normal detects changes in the mean and covariance matrix of a sequence of multivariate Gaussian random variables. It uses the log likelihood of empirical Gaussian distribution, for the computation of the cost function. Window AR estimates the least-squares estimates of the break dates obtained from a piecewise autoregressive (AR) model. NP-FOCuS is an online, non-parametric changepoint detection algorithm designed to efficiently detect distributional changes in real-time data streams by leveraging functional pruning techniques. For multivariate time-series, we instantiate n detectors and take the minimum detection time, where n is the number of features. CUSUM is a well-known QCD algorithm, which detects persistent shifts in the mean of a sequence. For multivariate time-series, we use the norm of the feature vector for the CUSUM chart. EWMA computes a running average of datapoints, placing exponentially decreasing weights on older observations.

Window L1, Window Normal, Window AR are implemented with `ruptures` (Truong et al., 2020), NP-FOCuS is implemented with `changepoint-online` (Romano et al., 2024b), and CUSUM and EWMA are implemented with `ocpdet` (Khamesi, 2022). We adapt window-based models from `ruptures`, originally designed for offline changepoint detection, to online QCD.

We fix both the window size and the burn-in interval at 30 frames, which is about three times smaller than the minimum maximum sequence length in our simulation experiments (100). It is also much smaller than the average lengths in the WISDM Actitracker. Most other hyperparameters remain at their default values; otherwise, they are specified in our code.

D.2. Preprocesses for WISDM Actitracker

Our preprocesses for the WISDM Actitracker dataset for both "labeled" and "unlabeled" subsets are comprised of the following procedures:

1. Extract user IDs ("user"), features ("X0", "X1", "X2", ..., "RESULTANT"), and frame labels (class").
2. Convert the string labels ("Jogging", "Walking", "Stairs", "Sitting", "Standing", "LyingDown") to binary numeric labels (0 for pre-change and 1 for post-change). In the "labeled" subset, "Sitting" is mapped to 0 and all other activities to 1. In the "unlabeled" subset, "Walking" and "Sitting" are mapped to 1, while the remaining activities are mapped to 0.
3. Split sequences that have a label transition from 1 to 0 to remove turn-back sequences.
4. Zero-pad outlier features (value $> 10^{12}$) (although some huge values such as 10^{10} still remain).

	User-labeled set	Machine-labeled set
#Sequences	83	51,326
#Frames	5,435	1,369,349
#Seqs. w/ 100% positive labels	37	76
#Seqs. w/ 0% positive labels	17	61
#Seqs. w/ mixed labels	29	51,189
Positive frame ratio	0.741	0.684
Mean length	65.5	26.7
Min length	1	1
Max length	565	54,401

Table 2. **Statistics of WISDM Actitracker after preprocesses.** A positive label here means that a frame (timestamp) is sampled from the post-change distribution. The user-labeled set and the machine-labeled set corresponds to the "labeled" set and the "unlabeled" set in the original WISDM Actitracker dataset. Note that the "unlabeled" set is actually labeled with a system developed in the WISDM Lab.

5. Normalize each feature to the range $[-1, 1]$.

6. Save feature vectors, labels, and changepoints to a file.

All the preprocesses are detailed in our code (`WISDMactitracker.ipynb`). Finally, we manually exclude the sequences of length 1 when evaluating QCD algorithms (in `calc_esARL_WISDM_cpmodels.py` and `calc_esADD_WISDM_cpmodels.py` of our code).

D.3. Statistics of WISDM Actitracker

Tab. 2 summarizes the statistics of the WISDM Actitracker dataset used in our experiments. The user-labeled and the machine-labeled subsets correspond to the "labeled" and the "unlabeled" subsets, respectively, in the original WISDM Actitracker dataset. Note that the "unlabeled" subset is, in fact, labeled with a system developed in the WISDM Lab (Kwapisz et al., 2011).

Fig. 18 shows histograms of sequence lengths of the user-labeled and machine-labeled subsets of the WISDM Actitracker dataset after the preprocesses, which are detailed in App. D.2. The sequence lengths exhibit substantial irregularity, and estimating the ARL and ADD is challenging.

Fig. 19 shows histograms of changepoint indices (timestamps) of the user-labeled and machine-labeled subsets of the WISDM Actitracker dataset after the preprocesses, which are detailed in App. D.2. They contain a variety of pre-change lengths, but the number of without-change sequences are limited, and estimating the ARL is challenging.

D.4. Numerical Computation of Variance

We describe the computation methods of the restricted variance of the detection time, $\hat{V} := \widehat{\mathrm{Var}}[\min\{\tau, a\}]$, used in our experiments.

The restricted variance of LB-ARL is defined as the naive empirical variance: $\hat{V} = \widehat{\mathrm{Var}}[\min\{\tau, T\}]$, which is consistent with the mean of LB-ARL ($\mathbb{E}[\min\{\tau, T\}]$). Given a dataset with size N, it is computed as $\frac{1}{N_{\mathrm{LB}}}\sum_{i=1}^{N_{\mathrm{LB}}}(\tau_i - \bar{\tau})^2$. $\bar{\tau}$ is the empirical LB-ARL, i.e., the detection time averaged over the subset in which $\tau_i \leq$ sequence length. The subset size is denoted by $N_{\mathrm{LB}}(\leq N)$.

There are several methods to numerically compute the variance of RMST (KM-ARL). We adopt `lifelines` (Davidson-Pilon, 2019), a standard Python library for survival analysis, to compute the KME and its variance. Specifically, we use `lifelines.utils.restricted_mean_survival_time` with *return_variance=True*, in which $\hat{V} = 2\int_0^a t\hat{S}(t)dt - (\int_0^a \hat{S}(t)dt)^2$, where $\hat{S}(t)$ denotes the empirical survival function, computed from the given dataset with size N. The derivation is given in the appendix in (Royston & Parmar, 2013), as is mentioned in the documentation[2].

[2] https://lifelines.readthedocs.io/en/latest/lifelines.utils.html#lifelines.utils.restricted_mean_survival_time.

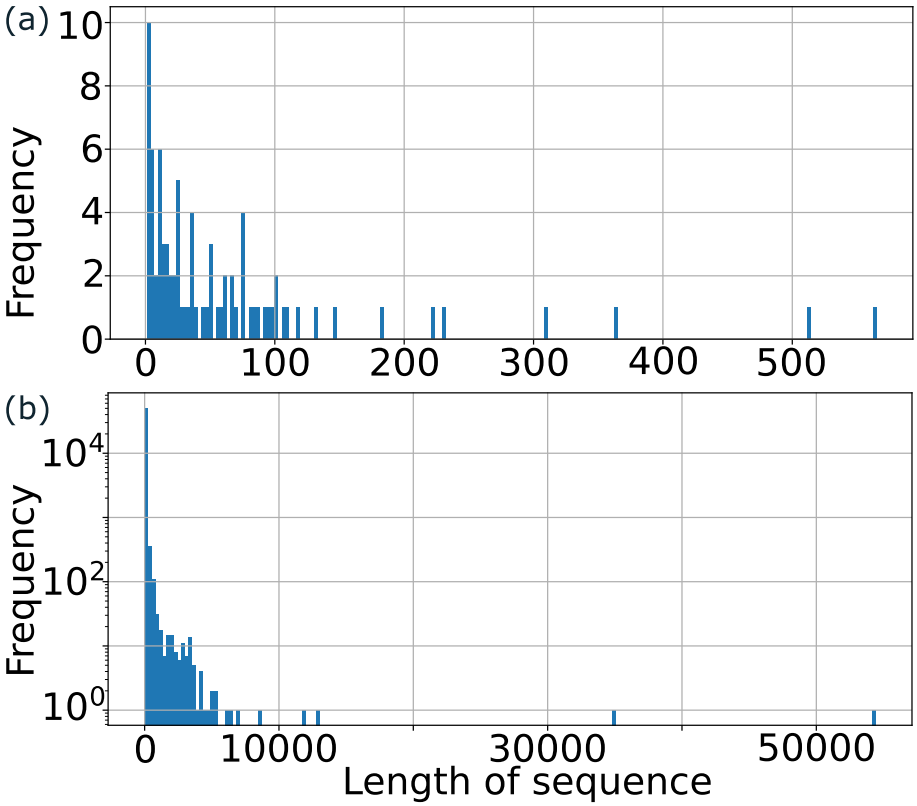


Figure 18. **Lengths of WISDM Actitracker sequences.** The sequence lengths exhibit substantial irregularity, and estimating the ARL and ADD is challenging. **(a) User-labeled subset (y-linear scale). (b) Machine-labeled subset (y-log scale).** The user-labeled and the machine-labeled subsets correspond to the "labeled" and the "unlabeled" subsets, respectively, in the original WISDM Actitracker dataset. Note that the "unlabeled" subset is, in fact, labeled with a system developed in the WISDM Lab (Kwapisz et al., 2011).

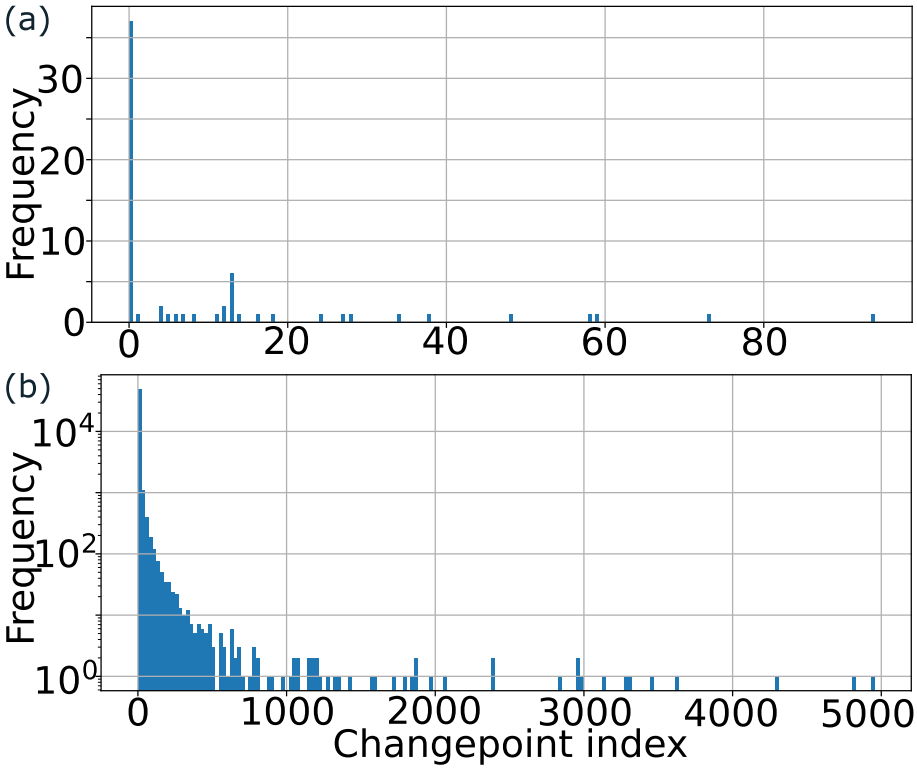


Figure 19. **Changepoint indices of WISDM Actitracker sequences.** Changepoint index here means the timestamp at which a change occurs, i.e., the pre-change length. They contain a variety of pre-change lengths, but the number of without-change sequences are limited, and estimating the ARL is challenging. **(a) User-labeled subset (y-linear scale).** The number of sequences without a changepoint is 17 out of 83. **(b) Machine-labeled subset (y-log scale).** The number of sequences without a changepoint is 61 out of 51326. The user-labeled and the machine-labeled subsets correspond to the "labeled" and the "unlabeled" subsets, respectively, in the original WISDM Actitracker dataset. Note that the "unlabeled" subset is, in fact, labeled with a system developed in the WISDM Lab (Kwapisz et al., 2011).

Why do KM-ARL/ADD reduce variance compared to LB-ARL/ADD? We elaborate on the discussion of our experimental results on the real-world dataset. In short, $\hat{V}$ of LB-ARL tends to larger than that of KM-ARL because the former can use only $N_{\mathrm{LB}} \leq N$, while the latter can exploit all N sequences. LB-ARL/ADD become unstable and exhibit high variance, particularly at large QCD thresholds, where N_{LB} tends to be small because the detector often fails to raise an alarm within the sequence length. In contrast, KM-ARL/ADD do not suffer from this issue because they are calculated from a constant number of sequences N, regardless of the threshold, which enhances their robustness against censoring.

E. Supplementary Related Work

QCD metrics. General QCD metircs include the Shiryaev's ARL (Shiryaev, 1963), Lorden's worst-case ADD (Lorden, 1971), Pollak's supremum conditional ADD (Pollak, 1985), exponential penalty detection delay (Poor, 1998), probability of false alarm (PFA) (Tartakovsky, 2019), precision-recall (Van den Burg & Williams, 2020), Jaccard index (Van den Burg & Williams, 2020), F_β score (Van den Burg & Williams, 2020), range-based precision-recall (Tatbul et al., 2018), NAB (Lavin & Ahmad, 2015), SoftED (Salles et al., 2024), Hausdorff distance (Van den Burg & Williams, 2020), adjusted Rand index (Hubert & Arabie, 1985), variation of information (Arabie & Boorman, 1973), and segmentation covering metric, (Everingham et al., 2010; Arbelaez et al., 2010). In this paper, we focus on the standard ARL and Bayesian ADD (expectation is taken over the changepoint) because they are intuitive for practical evaluation and are widely used in proving the optimality of many QCD algorithms (Tartakovsky, 2019).

To our knowledge, no existing metric explicitly addresses random irregular sequence lengths in QCD. Some metrics consider sequence truncation, such as the inverse of the false alarm rate over a finite time interval Δt (Alakent & Mutlu, 2018), which approximates the ARL for large Δt. The PFA in horizon (Huang & Veeravalli, 2024) is defined as the PFA measured on truncated sequences. However, these metrics are designed for fixed, finite-length sequences and introduce significant bias when applied to irregular-length sequences.

ARL and ADD estimation. There are many studies that derive the ARL and ADD of specific control charts or distributions: (Reynolds, 1975; Crowder, 1987; Saccucci & Lucas, 1990; Knoth & Knoth, 1998; Fu et al., 2002; Areepong & Peerajit, 2022; Haq & Woodall, 2022; Sunthornwat et al., 2023). Li et al. (2014) review simulation methods for computing the ARL and ADD of CUSUM (Page, 1954) and EWMA (Roberts, 1959), on simulation data. The references therein analyze Markov chain methods and integral equation methods. Alakent & Mutlu (2018) define the ARL of the Shewhart chart as the inverse of the alarm probability. In contrast, we propose the estimators of the ARL and ADD for arbitrary control charts, QCD models, and underlying distributions. Qiu (2013) defines a maximum length of the sequence for the ARL computation and excludes sequences that fail to detect changepoints by the maximum length. This estimator is similar to the LB-ARL and LB-ADD and exhibits a more substantial negative bias than our ARL estimator due to truncation, as proven in Sec.4 and demonstrated in Sec.5.

Is ARL always informative? (Mei, 2008) is the first paper in which the author systematically questions the appropriateness of the ARL. He claims that (i) a detection scheme with finite detection delay can have infinite ARL, where the detector can raise false alarms with probability 1, and (ii) under the standard minimax formulation with the ARL, we are in danger of finding a detection scheme that focuses on detecting larger changes instead of smaller changes. Claim (i) is theoretically relevant but practically irrelevant. It points out, in theory, that if the ARL is not assumed to be finite, one can construct a detector that raise false alarms. In practice, however, a detector with a small finite detection delay and an (approximately or numerically) infinite ARL is considered to perform well. This is not problematic in real-world applications. We agree with Claim (ii) because the minimax formulation of QCD, or more generally, the min and max operations are sensitive to outliers and fail to capture the average behavior of the model. For this reason, when used, they are typically complemented by average-based metrics such as the ARL and ADD. In summary, we encourage researchers to choose evaluation metrics with a clear understanding of their strengths and limitations.

QCD model for finite sequence length. To our knowledge, (Huang & Veeravalli, 2024) is the first and only work on the QCD problem on sequences of finite length. They propose a CUSUM variant for a fixed finite sequence length. If the model fails to detect the change before the horizon, they set the delay to $T - \nu$, causing a downward bias in the ADD.

QCD for survival analysis. There are many studies on the application of changepoint detection to survival analysis. In (Biswas & Kalbfleisch, 2008; Sego et al., 2009; GANDY et al., 2010; Phinikettos & Gandy, 2014; Gomon et al., 2024; Sasikumr & Sujatha, 2025), they propose variants of CUSUM to estimate survival time on right-censored data, which is orthogonal to our focus: accurately estimating ARLs and ADDs of arbitrary QCD algorithms on right-censored sequential data with irregular lengths. Gierz (2020) consider changepoint problems concerning a shift in a distribution for a set of time-ordered observations under censoring or truncation. Polunchenko (2016) derive a closed-form formula for the survival function of the generalized Shiryaev-Roberts (GSR) detection time under Brownian motion.

Parametric and closed-form approaches to ARL and ADD. The exponential approximation of the underlying distribution provides accurate estimation of ARL and ADD in some scenarios. Also, closed-form expressions for ARL are

known for some limited cases (e.g., (Fu et al., 2002; Areepong & Peerajit, 2022; Sunthornwat et al., 2023). However, it has been also known that: (1) the exponential approximation does not work well if the distribution deviates from exponential (Borror et al., 2003; Pehlivan & Testik, 2010); and (2) closed-form expressions are not always available—it depends on distributions and detectors. This is why we emphasize that our estimators are applicable to *arbitrary* underlying distributions (=non-parametric) and *arbitrary* QCD models, a key contributions that clearly distinguishes our work from prior approaches.

Supplement to main text. Ahmed et al. (2025) propose DeepLLR-CUSUM and uses KME and RMST to empirically evaluate it under *fixed* length truncation. Theoretical analysis of ARL and ADD estimations is not their primary focus.

F. Supplementary Discussion

Terminology. In general, the terms "average", "mean", and "expected" are used interchangeably when referring to the ARL and ADD. The ARL is also referred to as the ARL to false alarm (Tartakovsky, 2019) or the average time to signal (Li et al., 2014). The ADD with a fixed changepoint is also referred to as the conditional expected delay to detection (CEDD) (Tartakovsky, 2019). The ADD with taking expectation over the changepoint is referred to as the Bayesian ADD or, simply, ADD (Tartakovsky, 2019). In our paper, we refer the Bayesian ADD as the ADD (with $T = \infty$). In the statistical theory of control charts, including change detection in survival monitoring (or monitoring time-to-event data), the ARL and ADD are sometimes referred to as the in-control or zero-state ARL and the out-of-control or steady-state ARL, respectively (Saccucci & Lucas, 1990; Sasikumr & Sujatha, 2025; Lim & Lee, 2025). Strictly speaking, the steady-state ARL is defined as $\mathbb{E}[\tau \mid \nu = 0]$, not $\mathbb{E}[\tau - \nu \mid \tau \geq \nu, \nu < \infty]$.

How to define T and $T^*_{\max}$. We clarify how to define the distribution of the truncation length (denoted by F_T here), focusing on ARL, T (truncation length), and $T^*_{\max}$ (the least upper bound for the support of F_T, i.e., the maximum possible sequence length). A parallel discussion applies to ADD. In brief, given an evaluation dataset, we can arbitrarily define F_T. In practice, once a test dataset is collected for model evaluation, F_T may be taken as the empirical CDF of the test dataset (or a mollified (smoothed) version of it). Accordingly, $T^*_{\max}$ can be defined as the maximum sequence length in the test dataset, denoted by $T_{\max}$ in the main text.

Tighter bound. In Cor. 1.1 in (Stute, 1994), an additional tighter *upper* bound is available, which potentially can be used to derive tighter bounds for the KM-ARL and KM-ADD.

Theoretical computational efficiency. The computational complexity of KM-ARL is $\mathcal{O}(N) + \mathcal{O}(T_{\max})$ (plus event time sorting if necessary), where N is the number of sequences in the dataset and $T_{\max}$ is the max sequence length. Bucketing the sequences into $T_{\max}$ bins takes $\mathcal{O}(N)$, counting n_j^{ARL} and d_j^{ARL} takes $\mathcal{O}(T_{\max})$, computing $\hat{S}^{\mathrm{ARL}}$ takes $\mathcal{O}(T_{\max})$, and integrating the step-like function $\hat{S}^{\mathrm{ARL}}$ takes $\mathcal{O}(T_{\max})$.

Empirical computational efficiency. As mentioned in App. D, the computational costs of metric evaluation, including KM-ARL and KM-ADD, are negligible in our experiments compared with the time required to run the QCD algorithms. For significantly longer sequences (e.g., $\gtrsim 10^3$, which is our maximum length) and larger datasets (e.g., $\gtrsim 10^4$, which is our maximum size), we recommend splitting the dataset and aggregating the ARLs and ADDs, although it can degrade evaluation accuracy.

Evaluation on significantly small datasets. Evaluation of ARLs and ADDs on significantly small datasets such as the user-labeled subset of the WISDM Actitracker is challenging, as shown in Fig. 15 in App. C.5. In such cases, consider applying finite-sample bias correction methods (e.g., bootstrapping) or increasing samples.

Implementation in Python. We implement our estimators based on Python (Van Rossum & Drake, 2009) because it is more compatible with machine learning development than R (R Core Team, 2021).

Variance estimation. While our primary focus is on bias estimation, we also provide a rough estimate of variance. The variance of LB-ARL (naive, standard definition of variance) is given by

$$\int_0^a (t - \bar{t})^2 dF(t), \tag{112}$$

where a denotes the cut-off time, F the CDF of the event time, and $\bar{t}$ the mean event time. On the other hand, Akritas (2000) provides the asymptotic variance of the RMST—analogous to the variance of KM-ARL in our setting:

$$\int_0^a \frac{S(t)}{1-H(t)}(t-\bar{t})^2 dF(t) \tag{113}$$

(Eq. (4) in (Akritas, 2000)), where S denotes the survival function and H the CDF of $\min\{\text{event time}, \text{censoring}\}$. Compared with LB-ARL (112), the variance of KM-ARL (113) is computed as a weighted expectation of $(t-\bar{t})^2$, with its scale determined by the weight $\frac{S(t)}{1-H(t)}$. Because $S(t) \in [0,1]$, this weight can grow large when $1-H(t)$ is small unless $S(t)$ decreases sufficiently in the same region. Consequently, the variance integral in (113) can be dominated by the interval where t is large and close to the least upper bound of $H(t)$. Therefore, the variance of KM-ARL can exceed that of LB-ARL when both event time and censoring have heavy tails and a is large; otherwise, it may be smaller.

Finally, the assumptions and background theory in Akritas (2000) are intricate and require careful examination to validate this expression in QCD—one of the main challenges in developing our bias theory. A full treatment would warrant a separate paper.

F.1. More about Independent Censoring Assumption

We extend our discussion in Sec. 6 about potential alleviation of the independent censoring assumption.

F.1.1. Effect of Heavy, Dependent Censoring

Demonstrating the numerical effect of bias under violations of the independent censoring assumption would be helpful for practitioners. According to Fig. 2 in (Malmquist, 2025), under a constant hazard rate and with the number of subjects (sequences) = 150, it is reported that the mean squared error of the survival function over time is approximately given by:

- For independent censoring (uniform censoring):
  - < 0.01 for 10% censoring.
  - < 0.01 for 50% censoring.
  - < 0.01 for 90% censoring.
- For dependent censoring (occurring just before event time (detection)):
  - < 0.01 for 10% censoring.
  - ≈ 0.4 for 50% censoring.
  - ≈ 2.1 for 90% censoring.

This result shows that heavy, dependent censoring inflates the bias. It aligns with our statement in Sec. 6, where we clarify the appropriate use cases for KM-ARL and KM-ADD.

F.1.2. Solutions

Background: Restricted Mean Survival Time (RMST). To mitigate the bias arising from dependent censoring and finite time interval, we must quantify the effect of the bias on the survival function or its integral within a finite time interval, known as the restricted mean survival time (RMST) in survival analysis. It is the counterpart of KM-ARL and KM-ADD in survival analysis and is defined as the expected survival time up to a specified horizon (denoted by h here), equivalently the area under the survival curve from 0 to h. It has been known that the RMST is systematically biased under dependent censoring (Klein & Moeschberger, 1984; Rivest & Wells, 2001; Ebrahimi & Molefe, 2003).

Solutions. Many studies have developed mitigation techniques for bias arising from dependent censoring. The following works, among others, suggest promising directions for deriving KM-ARL/ADD under dependent (informative) censoring or alleviating the resulting bias.

- (Ebrahimi & Molefe, 2003): Develops a consistent estimator of the survival function under dependent censoring, which may extend to estimating ARL and ADD in QCD.

- (Hsu & Taylor, 2010): Proposes a robust weighted KME that uses baseline prognostic covariates to correct bias in the marginal survival function when censoring is dependent.
- (Lin et al., 2023): Applies inverse probability of censoring weighting (IPCW); at each time t, each subject (patient) still under observation is weighted by the inverse of their estimated probability of remaining uncensored up to t.
- (Crommen et al., 2025): Summarizes several bias-correction strategies:
  - Nonparametric marginal estimation under a known copula: Given a specified copula for (τ, C), observable probabilities uniquely determine the marginals. Step-function estimators are constructed that reduce to the KME under independence and yield a consistent estimator under dependent censoring.
  - Likelihood estimation in copula models: Corrects bias by modeling the joint distribution of (τ, C); parametric copulas enable joint likelihood maximization with identifiable parameters and consistent, asymptotically normal estimates.
  - Correction of regression functionals: Embeds dependent censoring in copula-based Cox and related semi-parametric models to adjust hazard ratios, covariate effects, and causal effects.
  - Machine-learning and partial-identification approaches: Addresses bias when flexible modeling or weaker assumptions are required, using both deep and non-deep learning methods.

F.2. Relevance of Requiring Datasets with Multiple Sequences with Changepoint Labels

In this paper, we focus on the situations where datasets of multiple sequences with changepoint labels are available. This problem setting is common in machine learning. Although there are many studies of QCD on unlabeled datasets or pure simulations, there are also many studies on labeled datasets.

There are a wide variety of labeled datasets with multiple sequences and labeled changepoints (or annotations that can be seen as changepoints).

- Human sensing: WISDM Actitracker (Kwapisz et al., 2011) includes both human- and machine-labeled changepoints. We use this real-world sensing data in our experiments.
- Twitter sentiment-change detection: Twitter Data Stream (US Airline Sentiment) (Makone, 2016), a collection of tweets annotated with the tags of positive/negative/neutral, is used to detect sentiment changes in social media data stream (Bouchikhi et al., 2019): e.g., a sudden increase in negative tweets after an event.
- Speaker change detection: Many speech datasets, such as DIHARD dataset (Ryant et al., 2018) and IITG-MV phase 3 dataset (Haris et al., 2012), have speaker or language annotations, which can be used for speaker change detection (Mishra & Prasanna, 2024).
- Temporal action localization: THUMOS14 (Jiang et al., 2014) and ActivityNet (Heilbron et al., 2015) are standard benchmark datasets in temporal action localization (Wang et al., 2023a). They contain a number of videos with temporal annotations of human actions (e.g., Horseback riding, Diving, Long jump, etc.), where timestamps with action changes can be seen as changepoints.
- Earthquake detection: STEAD (Mousavi et al., 2019) and several datasets in SeisBench (Woollam et al., 2022) offer datasets of seismic signals with annotation of event times.

Furthermore, we can create labeled datasets from unlabeled real-world dataset by injecting changepoints:

- Sound anomaly detection: Normal and abnormal audio data recorded from industry machine are concatenated to synthesize changepoints (Gopalan et al., 2021). The base dataset is the MIMI Dataset (Purohit et al., 2019).
- Social event detection: Based on the unlabeled MIT Cellphone Data (Eagle & Pentland, 2006), annotations are defined by linking changepoints to calender events (e.g., sponsor meeting, Presidents Day, spring break, etc.) (Kei et al., 2025).
- Social event detection: Similarly to (Kei et al., 2025), annotations are provided for Enron Email Data (Klimt & Yang, 2004) based on actual corporate events (e.g., Federal Energy Regulatory Commission's decisions, the CEO's public protests, the bankruptcy filing, etc.) (Wang et al., 2023b).

Of course, we can create labeled datasets via simulation (e.g., KDD Cup 1999 intrusion detection dataset (Stolfo et al., 1999)).

Thus, real-world applications include: human sensing and action recognition (Bishop, 2006; Wang et al., 2023a), sentiment-change detection on social media data stream (Bouchikhi et al., 2019), speaker/language change detection (Mishra & Prasanna, 2024), earthquake detection (Woollam et al., 2022), sound anomaly detection (Gopalan et al., 2021), event detection from cellphone data or email data (Kei et al., 2025; Wang et al., 2023b), and intrusion detection (Stolfo et al., 1999), among others.

In summary, although our estimators cannot be used for unlabeled datasets, as noted in the main text, they apply to many real-world tasks listed above, among others.

Our contributions to machine learning community. Finally, we would like to note that analyzing how we evaluate performance metrics has been of general interest in machine learning (Hernàndez-Orallo et al., 2004; Lachiche et al., 2005; 2006; Pardo-Fernández & Castro, 2025). Recent top-tier conferences themselves emphasize this by requiring careful reporting of metrics in their Author Instructions and Reviewer Instructions, underscoring that reproducible and interpretable evaluation is a first-class concern, not an afterthought. Our work advances more robust, interpretable evaluation of QCD models, enabling empirical, intuitive model selection, as stated in the Abstract and Experiment sections.