Survival Analysis with Limited Overlap and Censoring Distribution Shift

Meera Krishnamoorthy MEERAK@UMICH.EDU
Division of Computer Science and Engineering
University of Michigan
Ann Arbor, USA

Donna Tjandra DOTJANDR@UMICH.EDU
Division of Computer Science and Engineering
University of Michigan
Ann Arbor, USA

Divya Shanmugam DIVYAS@MIT.EDU
Department of Computer Science
Cornell Tech
New York, USA

Amanda E. Kowalski AEKOWALS@UMICH.EDU
Department of Economics
University of Michigan
Ann Arbor, USA

Jenna Wiens WIENSJ@UMICH.EDU
Division of Computer Science and Engineering
University of Michigan
Ann Arbor, USA

Abstract

Survival analysis methods are often used to predict the time until the onset of an event in settings when the true time-to-event (TTE) may be censored during training. Such approaches typically assume uncensored data are representative of censored data and that the probability of censoring conditioned on the covariates remains constant over time, *i.e.*, there is no censoring distribution shift. However, both assumptions can fail in practice when censoring results from interventions targeted to individuals with particular comorbidities or genetic markers (such as prophylactic surgery when predicting time to cancer onset, or scheduled cesarean delivery and induction when predicting time to spontaneous labor) and changes in clinical policies alter which individuals are targeted for these interventions over time. To address this, we propose a new approach, cluster-weighted inference of time-to-event (CWITE), that remains accurate when these assumptions do not hold. Unlike existing approaches that ignore times-to-censoring (TTC) or treat them only as a lower bound of the TTE, CWITE leverages the insight that a subset of censored individuals are likely censored close to their true TTEs, and uses a novel mechanism to learn from such individuals. On the task of predicting time to spontaneous labor using real-world data, CWITE improves TTE accuracy for individuals similar to censored training data (mean absolute error: 6.50 days, 95% CI: [5.55, 7.40] vs. 7.82 days, [6.82, 8.82]) while maintaining comparable performance for those similar to uncensored training data (6.50 days, [5.54, 7.61] vs. 6.63 days, [5.67,

7.69]). Our results demonstrate that incorporating more specific supervision from censored training data can significantly improve TTE predictions in settings with limited overlap and censoring distribution shift, challenges common in real-world clinical data. Code to implement CWITE and reproduce all experiments in the paper is available at https://github.com/MLD3/CWITE.

1. Introduction

Survival analysis methods are widely used to estimate time-to-events (TTEs). They leverage data from both uncensored individuals who experience the event of interest and censored individuals who are observed to be event free up to some time before the event occurs. These approaches typically assume sufficient overlap between the covariate distributions of censored and uncensored individuals in the training data (Powell, 1984; Robins and Rotnitzky, 1992; Lee et al., 2018; Wang and Sun, 2022). However, this assumption breaks down when censoring results from interventions that target individuals with particular comorbidities or genetic markers (*e.g.*, prophylactic surgeries when predicting time to cancer onset or scheduled c-sections and inductions when predicting time to spontaneous labor), which can lead to censored and uncensored individuals having markedly different covariates (Metcalfe et al., 2019; ACOG, 2021).

This challenge is less of a concern when the probability of censoring remains constant across time and environments. For instance, if individuals with specific covariates consistently receive preventive interventions, their TTEs are never observed, and accurate predictions of these TTEs have limited value. This assumption of stable censoring probabilities underlies many approaches for handling informative or dependent censoring (Gharari et al., 2023; Zhang et al., 2024). However, in practice, this assumption frequently breaks down, not only due to rare events like pandemics, but also more routine factors such as evolving guidelines (*e.g.*, efforts to reduce unnecessary c-sections (The Leapfrog Group, 2025) or increase elective inductions at 39 weeks (Carmichael and Snowden, 2019)), seasonal strain (*e.g.*, flu season (Axios News, 2025), holiday staffing shortages (VISTA Staffing Solutions, 2024)), and policy changes (*e.g.*, limitations on COVID-19 booster eligibility (Gambino and Press, 2025) and puberty blocker access (Mayo Clinic Staff, 2023)). Additionally, personal preferences play a role; some individuals, whose covariates suggest they should receive interventions, may decline them, making accurate TTE prediction essential for personalized care (Press, 2012; Rosenberg and Trevathan, 2018; Antoine and Young, 2020; Keelan et al., 2021). We refer to this challenge as *survival analysis with censoring distribution shift*: a setting where the relationship between the covariates and the probability of censoring can change over time and across environments, and the goal is to produce accurate TTE predictions for all individuals, regardless of their probability of being censored.

In this work, we show that existing survival analysis methods often struggle to predict accurate TTEs for individuals similar to censored training data when overlap is limited, and as a result, can perform poorly under censoring distribution shift that causes individuals likely to be censored during training to become uncensored at test time. Standard approaches only penalize predictions that precede observed times-to-censoring (TTCs), which can lead to systematic overestimation (Powell, 1984; Lee et al., 2018; Wang and Sun, 2022). Inverse probability of censoring weighting (IPCW) avoids this by relying solely on reweighted uncensored training data, but fails when such data are not representative of

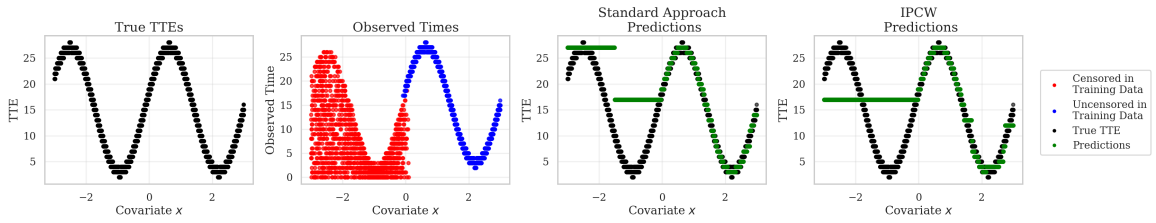

Figure 1: Failure modes of existing survival analysis methods derived from one of our synthetic data settings. (1) Ground-truth TTEs as a function of covariates. (2) Observed times for uncensored (blue) and censored (red) data. (3) Overestimation bias in standard methods. (4) Inaccuracy of IPCW under limited overlap.

censored training data (Robins and Rotnitzky, 1992; Dong et al., 2020). In summary, in such challenging settings, models receive either absent (in the case of IPCW) or unspecific (in the case of standard approaches) supervision for censored training data, resulting in poor generalization to individuals similar to these data under censoring distribution shift (**Figure 1**).

In light of these limitations, we propose an approach, Cluster-Weighted Inference of Time-to-Event (CWITE), that specifically leverages censored data during training. CWITE is motivated by two observations. First, given sufficiently informative input variables, individuals can often be grouped into clusters where members tend to exhibit similar TTEs. Second, preventive interventions are often triggered by clinical information that is also predictive of the event time itself. For example, when predicting time to spontaneous labor, pregnancy complications and evolving clinical trajectories may both indicate when labor is likely to occur and prompt induction or cesarean delivery before spontaneous labor begins (ACOG, 2021). Consequently, TTCs are not necessarily arbitrary: for some censored individuals, the timing of intervention may contain information about the underlying trajectory toward the event. At the same time, intervention timing is only partly determined by clinical factors, as hospital scheduling constraints, resource availability, institutional workflows, and patient preferences introduce additional variation (Knight et al., 2016; Coates et al., 2021). Censored individuals are therefore not always observed at a fixed or uniformly distant time before their true TTEs; some may be censored close to when the event would otherwise have occurred. This is particularly plausible in early-event regions, where the range of possible TTC–TTE gaps is smaller and accurate prediction may be most clinically actionable.

Our approach identifies individuals likely censored close to their true TTE by clustering data in the input space. Within each cluster, the individual with the highest TTC is most likely censored closest to their true TTE, since higher TTCs leave less room for the gap between TTC and TTE. CWITE uses this cluster maximum as a reference point, increasingly penalizing predictions that deviate from a censored individual's TTC as that TTC approaches the cluster maximum.

We provide a theoretical analysis of the conditions under which CWITE performs optimally and propose adjustments to improve performance when these conditions are not satisfied. We then empirically compare CWITE to several baselines on synthetic, semi-synthetic, and real datasets with limited overlap between uncensored and censored training examples. On these datasets, CWITE produces more accurate TTE estimates for individuals resembling censored training data (without sacrificing accuracy for those similar to uncensored training data) compared to existing methods. The contributions of this work are as follows:

- We introduce a new survival analysis method, CWITE, designed to perform well under limited overlap and censoring distribution shift.
- We present a theoretical analysis outlining the conditions under which CWITE performs optimally, and describe adjustments to improve performance when these conditions are not met.
- We conduct empirical evaluations of CWITE on synthetic, semi-synthetic, and real datasets with limited overlap, demonstrating that CWITE yields more accurate TTE estimates across varying censoring probabilities compared to existing baselines.

Generalizable Insights about Machine Learning in the Context of Healthcare

We consider settings where censoring occurs via a preventive treatment, and the treatment decision is subject to factors that shift over time and across environments (e.g., clinical guidelines, societal attitudes, and scheduling constraints (Antoine and Young, 2020; Rosenberg and Trevathan, 2018; Press, 2012)). This paper offers two insights relevant to this setting. First, when the probability of censoring shifts across time and environments, a challenge we call censoring distribution shift, existing survival analysis methods can produce inaccurate predictions for individuals who look similar to those censored at training time, but who are uncensored at test time. Stratifying test-time evaluation by similarity to censored vs. uncensored training data is essential for diagnosing these failures, as aggregate metrics can mask substantial performance differences across these subgroups. Second, clustering individuals by covariates offers a practical mechanism for identifying which censored individuals are most likely censored near their true event times.

2. Problem Formalization & Background

In survival analysis, each individual i is associated with covariates $\mathbf{x}^{(i)} \in \mathbb{R}^d$, event indicator $e^{(i)} \in \{0, 1\}$, and observed time $t^{(i)}_{\text{obs}} \in \mathbb{N}$. The observed time is the individual's TTE $t^{(i)}_E$ if $e^{(i)} = 1$ or TTC $t^{(i)}_C$ if $e^{(i)} = 0$. Like previous work, we assume right censoring and that the TTE, TTC, and censoring models can be identified from observed covariates (see Appendix Section 1 for details) (Lee et al., 2018; Wang and Sun, 2022; Gharari et al., 2023; Zhang et al., 2024).

Unlike prior work, we relax two common assumptions: (1) sufficient overlap in the probability of being uncensored ($P_{\text{train}}(e = 1 \mid \mathbf{x})$) and (2) stable probabilities of censoring between the training and test sets. First, we consider an asymmetric form of limited overlap, *limited uncensored training support*, where the probability of being uncensored for some

individuals in the training set, $P_{\text{train}}(e = 1 \mid \mathbf{x})$, is low. In contrast to causal inference, which defines overlap in terms of the probability of receiving treatment (i.e., the propensity score) (Zhou et al., 2020), we define overlap in terms of the probability of being uncensored. While both low and high values of the propensity score are problematic in causal inference, since they indicate poor representation of either the treated or untreated group, in survival analysis only low values of $P_{\text{train}}(e = 1 \mid \mathbf{x})$ are problematic, as uncensored individuals provide full supervision. High values do not pose an issue – in fact, when $P_{\text{train}}(e = 1 \mid \mathbf{x}) = 1$ for all individuals, survival analysis reduces to a standard supervised learning problem.

Second, we allow for censoring distribution shift, where $P_{\text{test}}(e = 1 \mid \mathbf{x}) > P_{\text{train}}(e = 1 \mid \mathbf{x})$. This makes it challenging to predict TTEs for individuals with little or no uncensored training support, *i.e.*, those rarely observed to have an event but requiring reliable estimates at test time. Such shifts can occur, for example, when policy changes (*e.g.*, fewer prophylactic surgeries), changing resource constraints, or patient preferences lead patients who would have previously been censored to become uncensored. Throughout, we use limited overlap and censoring distribution shift to refer to settings with low uncensored support in training and high uncensored support at test time.

Our goal is to produce accurate TTE estimates for all individuals, including those with a low probability of being uncensored during training. With this goal in mind, in the next section, we review existing survival analysis methods and highlight their limitations under limited overlap and censoring distribution shift. Our focus is not on model architecture, but on how methods incorporate censored training data into their objectives to predict TTEs.

2.1. Existing Approaches in Survival Analysis

Standard survival analysis methods typically assume uninformative or randomly occurring censoring, implying no censoring distribution shift. These methods learn $\hat{t}_E$ by minimizing different loss functions. One class of approaches, similar to Powell's censored least absolute deviation (CLAD) estimator (Powell, 1984), uses a loss that penalizes predicted TTEs for censored individuals only when the prediction falls *before* their observed TTC. In contrast, another class of approaches—including many deep learning-based survival analysis methods (Lee et al., 2018; Wang and Sun, 2022; Pakbin et al., 2021; Lee et al., 2020)—explicitly encourages predictions for censored individuals to fall *after* their TTCs by maximizing the predicted survival probability beyond the TTCs. Both losses can yield biased estimates for test individuals in covariate regions with limited uncensored training data. Because overestimation of censored individuals' TTEs is not penalized, predictions in these regions may be systematically inflated (**Figure 1**).

Inverse probability of censoring weighting (IPCW) (Robins and Rotnitzky, 1992; Dong et al., 2020) attempts to correct for this by reweighting uncensored examples to approximate conditional independence between censoring and the TTE. This reweighting helps mitigate censoring distribution shift in settings where uncensored support is limited in training but more prevalent in testing. Unlike standard approaches, IPCW adjusts for underrepresented strata by amplifying their influence during training, improving generalization in those regions. Still, IPCW depends on some degree of uncensored support in training. When that support is sparse or absent, IPCW may yield high-variance or biased estimates. This limitation is well documented in analogous settings in the causal inference literature (where

overlap is defined with respect to treatment rather than censoring), where techniques like weight clipping can reduce variance but cannot eliminate bias from a lack of support (Li et al., 2018; Zhou et al., 2020; Thomas et al., 2020; Imbens and Rubin, 2015).

In summary, both standard and IPCW methods break down when uncensored training support is absent; standard methods are particularly vulnerable, as they lack any mechanism to address the absence.

3. Proposed Approach

Our approach, CWITE, consists of two stages. In the first stage, we identify censored individuals likely to provide more informative supervision. In the second stage, we optimize a loss function that upweights censored individuals deemed more informative. The pseudocode for CWITE is in Appendix Algorithm 1.

Stage 1: Clustering to Identify Informative Censored Cases. We begin by clustering the covariates of censored and uncensored training data and identifying the maximum TTC in each cluster. Any clustering approach may be used, provided it satisfies Assumption 1, defined in the next section.

Let censored individual i belong to cluster k, *i.e.*, $i \in C_k$, where C_k denotes the set of all individuals assigned to cluster k. Let $C_k^c \subseteq C_k$ denote the subset of censored individuals in cluster k. For each censored individual i, we denote the maximum TTC among censored individuals in their cluster: $t_{\max}^{(i)} = \max_{j \in C_k^c} t_{\text{obs}}^{(j)}$. Out of all individuals in cluster k, the individual for whom this value corresponds to their TTC is likely censored closest to their true TTE (formally justified based on Assumption 1 in the next section), and thus represents the most reliable censored point in the cluster.

Stage 2: Novel Weighting Scheme for TTE Prediction. We train a model to predict TTEs using censored and uncensored training data. For uncensored individuals, we minimize the IPCW loss to correct for non-random censoring. For censored individuals, we introduce a novel loss formulation. For each censored individual i, we minimize a loss that encourages the model's predicted TTE to be close to their TTC, thereby discouraging overestimation. To avoid significant underestimation, we scale each censored individual's loss contribution by the ratio $g(t_{\text{obs}}^{(i)})/g(t_{\max}^{(i)})$ where $g(\cdot)$ is a monotonic weighting function (*e.g.*, $g(x) = e^x$). This ratio acts as a *soft filter*, assigning negligible weight to censored data whose TTCs are far from the cluster maximum, and emphasizing those likely to have been censored near their true TTE. In well-formed clusters, where members share similar underlying TTEs, this weighting prioritizes censored individuals whose TTCs are more informative.

To balance the influence of censored and uncensored losses, we include a tunable hyperparameter λ. Thus, given that θ denotes the parameters learned by CWITE, the complete objective minimized by CWITE is:

$$\min_{\theta} \left(-\sum_i e^{(i)} \left(\frac{1}{P\left(e^{(i)} \mid \mathbf{x}^{(i)}\right)} \log P\left(\hat{t}_E^{(i)} = t_{\text{obs}}^{(i)} \mid \mathbf{x}^{(i)}; \theta\right) \right) - \lambda \sum_i \left(1 - e^{(i)}\right) \left(\frac{g\left(t_{\text{obs}}^{(i)}\right)}{g\left(t_{\text{max}}^{(i)}\right)} \log P\left(\hat{t}_E^{(i)} = t_{\text{obs}}^{(i)} \mid \mathbf{x}^{(i)}; \theta\right) \right) \right)$$

We describe practical considerations for selecting g and λ in the Experimental Setup Section, Results Section, and Appendix Section 5.

3.1. Theoretical Analysis

Unlike existing approaches to survival analysis, CWITE does not require overlap between censored and uncensored individuals or a constant censoring mechanism across environments. Instead, CWITE relies on a few structural assumptions about the training data. We formalize these assumptions below and analyze how violations of these assumptions influence the effectiveness of CWITE. Along with the C_k^c notation defined above, we use the following notation throughout: $\Delta^{(i)} = t_E^{(i)} - t_{\text{obs}}^{(i)}$ is the time between a censored individual's TTE and their TTC.

Assumption 1 (Within-Cluster TTE Similarity) *We assume that the clustering procedure groups individuals such that within each cluster k, true TTEs are relatively homogeneous. In contrast, variation in TTCs among censored individuals may reflect additional censoring noise (e.g., due to administrative factors) that inflates their variance beyond that of the underlying TTEs. Formally, we assume that* $\text{Var}_{i \in C_k^c}(t_C^{(i)}) > \text{Var}_{i \in C_k^c}(t_E^{(i)})$.

When the above inequality holds, the excess variance in TTCs is more likely due to censoring noise than true variation in TTEs. As a result, within a cluster, censored individuals with larger TTCs (closer to t_{max}) are likely to be closer to their true TTEs.

Given that we assume identifiability of TTEs and censoring from observed covariates alone, Assumption 1 is more likely to hold in settings with sufficiently large sample sizes and low-to-moderate dimensionality, where meaningful clustering based on covariates is feasible. In contrast, in small sample or high-dimensionality settings, the assumption becomes less plausible, as noise dominates and clusters may reflect irrelevant or weakly predictive features. However, we hypothesize that CWITE can remain effective even when Assumption 1 is weakened, provided the weighting function $g(\cdot)$ is chosen appropriately. We empirically test this in the Results Section.

Assumption 2 (Within Cluster Close-to-TTEs TTCs) *We assume that in each cluster, at least one censored individual is observed sufficiently close to their true TTE. Formally, for some small threshold $\epsilon > 0$ and for all k, we assume:* $P\left(\min_{i \in C_k^c} \Delta^{(i)} \leq \epsilon\right) \gg 0$.

Below, we show two results. First, the probability that at least one censored individual's TTC is close to their TTE is monotonically increasing in $|C_k^c|$. Second, while the exact dependence on $p = \min_{i \in C_k^c} P(\Delta^{(i)} \leq \epsilon)$ is harder to characterize directly, a lower bound on

this probability $\left(1-(1-p)^{|C_k^c|}\right)$ is monotonically increasing in p, indicating that as each individual's probability of being censored near their TTE increases, Assumption 2 becomes more likely to hold.

Claim 1 *The probability that at least one censored individual in cluster k is censored within ϵ of their TTE, $P\left(\min_{i\in C_k^c}\Delta^{(i)}\leq\epsilon\right)$, is monotonically increasing in $|C_k^c|$, and its lower bound $1-(1-p)^{|C_k^c|}$ is monotonically increasing in p.*

Proof 1 *Assuming independence across individuals in C_k^c, we have: $P\left(\min_{i\in C_k^c}\Delta^{(i)}\leq\epsilon\right)=1-\prod_{i\in C_k^c}\left(1-P(\Delta^{(i)}\leq\epsilon)\right)$. We prove that the above function monotonically increases with respect to $|C_k^c|$ in the following way: adding individual j to the cluster multiplies the product $\prod_{i\in C_k^c}\left(1-P(\Delta^{(i)}\leq\epsilon)\right)$ by $(1-P(\Delta^{(j)}\leq\epsilon))$, which lies in $(0,1)$. Therefore, it can only decrease the product, meaning that it can only increase $1-\prod(\cdot)$. We prove that the above function's lower bound monotonically increases with respect to p in the following way: since $p=\min_{i\in C_k^c}P(\Delta^{(i)}\leq\epsilon)$, we can bound $\prod_{i\in C_k^c}\left(1-P(\Delta^{(i)}\leq\epsilon)\right)\leq(1-p)^{|C_k^c|}$ which implies $P\left(\min_{i\in C_k^c}\Delta^{(i)}\leq\epsilon\right)\geq 1-(1-p)^{|C_k^c|}$. This lower bound is monotonically increasing in p.*

This implies that Assumption 2 is more likely to hold in clusters with more censored individuals. Additionally, as the minimum individual probability of being censored near one's true TTE increases, the probability that Assumption 2 holds also increases.

Unlike Assumption 1, we do not introduce techniques to explicitly mitigate violations of Assumption 2. Instead, we treat Assumption 2 as a descriptive property of the data: an indicator of when CWITE is likely to perform well or poorly. In real survival analysis datasets, this assumption cannot be directly verified because true TTEs are unknown for censored individuals. However, its plausibility can be assessed using domain knowledge about whether a set of features increases both event risk and prompts intervention. When this is true, censored individuals may have TTCs close to when the event would otherwise have occurred. This is especially plausible for early true TTEs, where possible TTC–TTE gaps are smaller and accurate prediction is often most clinically important. In semi-synthetic datasets, Assumption 2 can be tested directly using the uncensored individuals used to generate labels (Qi et al., 2023).

Following prior work (Qi et al., 2023), we construct such datasets using a generative model for TTEs and TTCs. To this end, we introduce a model for the gap between TTE and TTC, $\Delta^{(i)}=t_E^{(i)}-t_C^{(i)}$.

Assumption 3 (Model for TTE-TTC Gap) *To avoid introducing unobserved confounding between $t_E^{(i)}$ and $t_C^{(i)}$ (details in Appendix Section 3), we assume that $t_E^{(i)}$ and $t_C^{(i)}$ are generated from a shared latent function $f(\mathbf{x}^{(i)})$ with independent uniform deviations. We assume that TTEs are largely covariate-determined with small additive noise: $t_E^{(i)}=f(\mathbf{x}^{(i)})+\alpha^{(i)},\quad\alpha^{(i)}\sim\mathcal{U}(0,\eta)$, η small. TTCs are drawn with greater heterogeneity: $t_C^{(i)}=f(\mathbf{x}^{(i)})-\beta^{(i)},\quad\beta^{(i)}\sim\mathcal{U}(0,f(\mathbf{x}^{(i)}))$. Under this model, the gap between censoring and the true TTE, $\Delta^{(i)}=\alpha^{(i)}+\beta^{(i)}$, follows a trapezoidal distribution with lower bound 0, upper bound $\eta+f(\mathbf{x}^{(i)})$, and a flat region between η and $f(\mathbf{x}^{(i)})$.*

This assumption is grounded in domain knowledge and aligns with prior work. Typically, the relationship between covariates and TTEs is considered predictable, whereas the relationship between covariates and TTCs can exhibit substantially greater heterogeneity, particularly when the probability of being censored depends on deterministically administered interventions due to administrative, personal, and institutional factors (Knight et al., 2016; Coates et al., 2021). Furthermore, from a statistical perspective, the quantiles of any distribution are uniformly distributed. Past work has leveraged this fact to motivate modeling latent variables as uniformly distributed, even when not estimating quantiles directly, as a flexible way to capture heterogeneity without imposing strong parametric constraints (Bjorklund and Moffitt, 1987; Heckman and Vytlacil, 1999; Kowalski, 2023). We use this assumption to define synthetic and semi-synthetic DGPs without unobserved confounding between TTEs and TTCs; CWITE itself does not require this exact DGP at deployment.

Under Assumption 3, for small ϵ, $P(\Delta^{(i)} \leq \epsilon) = \frac{\epsilon^2}{2\,\eta\, f(\mathbf{x}^{(i)})} \propto \frac{1}{\eta f(\mathbf{x}^{(i)})}$ which implies that $p = \min_{i \in C_k^c} P(\Delta^{(i)} \leq \epsilon) \propto \frac{1}{\eta \max_{i \in C_k^c} f(\mathbf{x}^{(i)})}$. Thus, p increases as η and the maximum $f(\mathbf{x}^{(i)})$—and consequently the largest $t_E^{(i)}$—among censored individuals in cluster k decrease. As a result, in limited data settings (e.g., when $|C_k^c|$ is small), CWITE is expected to be most accurate for censored individuals with earlier TTEs.

4. Experimental Setup

We evaluate CWITE on synthetic, semi-synthetic, and real-world settings, with progressively decreasing control over the data-generating process from synthetic to semi-synthetic to real. Accordingly, the extent to which Assumptions 1 and 2 are satisfied likely diminishes across these settings.

In both synthetic and semi-synthetic settings, we generate TTEs and TTCs following Assumption 3, setting η to 1–10% of the maximum value of f, following past work simulating error in kernel-based learning and control systems (Maddalena et al., 2021). Censoring is then introduced via a covariate-driven censoring indicator function $e(\mathbf{x})$. To simulate censoring distribution shift, we apply this censoring mechanism only to the training and validation sets; in the test sets, all individuals are uncensored.

4.1. Synthetic Data

In the synthetic setting, we fully control the data-generating process. We generate two sets of 1,000 datasets each, totaling 2,000. In the first set, $\eta \in \{1, 2, 5\}$, ensuring that both Assumptions 1 and 2 hold (confirmed in Appendix Section 5). In the second set, $\eta \in \{5, 10\}$, which increases the gap between TTEs and TTCs, making Assumption 2 less likely to hold. Each dataset is defined by a fixed but unique $f(\mathbf{x})$, $e(\mathbf{x})$, and η, and contains $n = 10{,}000$ samples. See Appendix Section 4 for details. To isolate violations of Assumption 1, we reuse the first set of 1,000 datasets but randomly reassign 25% of individuals' cluster memberships after k-means clustering, directly degrading within-cluster TTE homogeneity without changing the data-generating process. We evaluate results from the second set stratified by true TTE quintile, since Assumption 2 is more likely to hold for individuals with earlier TTEs.

4.2. Semi-Synthetic Data

In the semi-synthetic setting, we have moderate control over the data-generating process. We construct a single dataset using real covariates $\mathbf{x}^{(i)} \in \mathbb{R}^{31}$ from the SUPPORT dataset (Knaus et al., 1995) (motivation for selecting this dataset in Appendix Section 4). This limits our control over feature distributions and sample size ($N = 6201$). The function $f(\mathbf{x}^{(i)})$ is defined as a neural network trained to learn the relationship between observed covariates and true TTEs of uncensored individuals in the SUPPORT dataset. The censoring mechanism $e(\mathbf{x})$ is generated based on clinical severity markers (coma hours, bilirubin, creatinine, mean blood pressure, and albumin) that are used in SOFA/APACHE II scores. Due to the input dimensionality and limited sample size, in this setting, Assumption 1 does not hold, while Assumption 2 holds for individuals with earlier true TTEs but not for those with later true TTEs (verified in Appendix Section 5). See Appendix Section 4 for details.

4.3. Real Data

In the real-data setting, we have the least control: covariates, TTEs, and censoring mechanisms are not simulated – they are from Michigan Medicine's electronic health record – and true TTEs for censored individuals are unknown. Although we cannot verify assumptions directly, the high dimensionality of the real-data covariate space can degrade distance-based clustering and make Assumption 1 less plausible. To reduce this risk, all real-data experiments used the same supervised feature-selection procedure. We ranked features by their association with TTE among uncensored training individuals and allowed each method to use the top 50, 100, or 200 features, or all standardized features, with the final feature set selected by validation performance. On the other hand, Assumption 2 is more likely to hold across a broader range of TTEs than in the semi-synthetic setting, owing to the larger sample size (45,632 vs. 6,201).

Our task is to predict time to spontaneous labor, defined by the onset of labor without medical induction (Patterson et al., 2008). Accurate prediction can benefit both patients and hospital systems (Arnold, 2022). Past work has stated that this prediction task can be characterized as having limited overlap and censoring distribution shift (ACOG, 2021; Antoine and Young, 2020). This study was approved by the University of Michigan institutional review board.

Covariates $\mathbf{x}$ include the most recent vitals and diagnoses of pregnant individuals at Michigan Medicine prior to 36 weeks gestation, as well as demographics and pregnancy-specific features. Individuals are labeled as spontaneous labor cases ($e = 1$) if a record of them going into labor precedes a record of them receiving an induction; TTE is the first recorded timestamp of them going into labor. All others are censored ($e = 0$), with observed time set to the time of admission prior to delivery. See Appendix Section 4 for details.

4.4. Baselines

We compare CWITE to three baselines introduced in the Problem Formalization & Background Section: "IPCW" (Robins and Rotnitzky, 1992; Dong et al., 2020); "Standard Margin," which assigns a non-zero loss to censored training samples only when their predicted TTE falls below their TTC (Powell, 1984); and "Standard," which explicitly encourages the model to predict TTEs greater than the TTC for censored individuals, a strategy widely

adopted in many deep learning-based survival analysis methods (Lee et al., 2018; Wang and Sun, 2022). All methods use the same multi-layer-perceptron-based feature extractor, allowing us to isolate the effect of the objective function under limited overlap and censoring distribution shift. Implementation details of our approach and baselines are in Appendix Section 4.

4.5. Evaluation Metrics

Survival analysis methods are typically evaluated using the concordance index (c-index), which assesses whether predicted TTEs correctly rank uncensored individuals relative to others' TTEs or TTCs (Hartman et al., 2023). However, the c-index ignores the magnitude of errors, limiting interpretability (Hartman et al., 2023). Because CWITE is designed to improve absolute TTE prediction, we use mean absolute error (MAE) as the primary metric; MAE expresses error directly in time units and avoids penalizing irrelevant ranking differences (Qi et al., 2023).

The MAE is computed only for uncensored individuals, which is appropriate in our synthetic and semi-synthetic settings where all test data are uncensored. In these settings, we label test points as censored-like or uncensored-like based on whether they would have been censored under the generated training rule e. In the real-data setting, we stratify test individuals by their estimated probability of being uncensored during training, $P_{\text{train}}(e_{\text{test}} = 1 \mid \mathbf{x}_{\text{test}})$, estimated using a logistic regression model fit on the training data. Because our claims concern behavior in regions where uncensored training support is effectively absent versus abundant, we first take all test individuals with an estimated propensity below 0.10, of whom 103 are uncensored, and then form a comparison group of equal size at the opposite extreme: the 103 uncensored test individuals with the highest estimated propensity (all at or above 0.78). Because MAE is computed only for uncensored individuals, matching the two subgroups on the number of uncensored individuals gives them equal precision. The propensity model's performance is evaluated in Appendix Section 5. Although censoring-aware metrics such as IPCW-MAE would be preferable in principle, they are not reliable here because the required propensity-score overlap is violated.

Along with reporting MAE in the main text, we report the c-index results in Appendix Section 5 to assess ranking performance, and additionally report overprediction rate in the real-data setting to assess potential relevance for hospital scheduling and resource planning. For c-index calculations, tied predictions receive half credit, and subgroup c-index values are computed using comparable pairs in which one member belongs to the subgroup and the other may be any individual in the test set.

Units for MAE. In the synthetic and real data, model outputs are in days, so MAE is reported directly in days. In semi-synthetic data, TTEs are discretized into 50 bins of equal size; MAE is computed in bins and converted to days.

Error bars and significance. For synthetic data, we evaluate MAE across 1,000 randomly generated datasets and report the median and empirical 95% CI. For each semi-synthetic task, we train 1,000 models with different splits and report the median and 95% CI. For real data, we fix one train/validation split, bootstrap test predictions 1,000 times, and report the median and 95% CI. Significance is assessed using a one-sided resampling test (Efron and Tibshirani, 1993).

5. Results

5.1. When Assumptions 1 and 2 Hold

In synthetic data where both assumptions hold (verified in Appendix Section 5), CWITE significantly outperforms baselines on individuals who would have been censored at training time (median MAE: 1.89, 95% CI: [0.74, 6.27] vs. 7.01, 95% CI: [1.29, 12.93]) while matching baseline accuracy on individuals who would have remained uncensored at training time (0.54, 95% CI: [0.11, 1.87] vs. 0.54, 95% CI: [0.08, 2.82]) (**Table 1**). It also significantly outperforms ablations of CWITE that do not cluster (6.49, 95% CI: [2.91, 11.43]) or weight censored training data differently based on the difference between the data's TTC and the cluster maximum's TTC (6.18, 95% CI: [1.68, 11.87]) . An ablation that assigns the maximum TTC of each cluster to all cluster members performs similarly (1.57, 95% CI: [0.46, 9.50]) but with a wider CI, indicating less stability. CWITE still does not match its own uncensored-like accuracy (0.54 vs. 1.89), since true TTEs are unavailable for censored individuals during training.

5.2. When Only Assumption 1 Is Violated

We next evaluate CWITE in two settings where Assumption 1 is violated but Assumption 2 still holds (verified in Appendix Section 5): (1) by randomly perturbing cluster assignments in synthetic data, and (2) by using semi-synthetic data with early TTEs ($<$ the median true TTE in the data, 816 days), where smaller sample sizes and higher dimensionality degrade within-cluster TTE homogeneity while Assumption 2 still holds. We implemented CWITE using k-means in both settings. Although k-means clustering degraded in the higher-dimensional semi-synthetic setting compared to the synthetic setting, it degraded less than the other clustering-based methods including k-means after PCA, Gaussian mixture models, and spectral clustering (**Appendix Table 13**).

In these settings, on individuals who would have been censored at training time, CWITE significantly outperforms all baselines. Its best performance occurs with a linear rather than exponential weighting function $g(\cdot)$. A linear weighting function spreads weight more evenly across censored points while still prioritizing those near the cluster maximum. By contrast, exponential weighting over-emphasizes points near the maximum, and with noisier clusters this amplifies errors. Specifically, in the synthetic data with the randomly perturbed cluster assignments, CWITE with linear weighting achieves a median MAE of 3.64 (95% CI: [1.62, 7.43]) and with exponential weighting 6.02 (95% CI: [1.82, 11.10]), both improving over

Table 1: Median MAE and 95% CIs for test individuals that are most similar to censored and uncensored training data in the **synthetic-data setting where Assumption 1 and 2 hold**.

Method	Censored-like Test Individuals	Uncensored-like Test Individuals
IPCW	8.29 [2.96, 13.83]	**0.54 [0.08, 2.82]**
Standard	8.89 [3.74, 14.31]	0.58 [0.15, 8.31]
Standard Margin	7.01 [1.29, 12.93]	0.56 [0.12, 8.54]
CWITE	**1.89 [0.74, 6.27]**	**0.54 [0.11, 1.87]**

Table 2: Median MAE (in days) with 95% CIs for test individuals in the **semi-synthetic data setting**. Assumption 1 is always violated. Results are segmented by whether Assumption 2 holds (true TTE <816 days) or not (true TTE ≥816 days), as verified in Appendix Section 5.

	True TTE <816 days (Assumption 2 holds)		**True TTE ≥816 days** (Assumption 2 violated)	
Method	Censored-like Test Individuals	Uncensored-like Test Individuals	Censored-like Test Individuals	Uncensored-like Test Individuals
IPCW	639.88 [543.41, 760.28]	335.28 [255.08, 472.03]	394.82 [352.42, 456.32]	350.54 [291.99, 423.92]
Standard	1328.66 [1224.20, 1427.31]	482.55 [374.73, 586.45]	421.41 [373.72, 467.26]	**299.77 [245.30, 355.72]**
Standard Margin	834.14 [300.37, 1329.83]	608.77 [205.08, 1266.22]	510.35 [258.48, 1077.61]	499.12 [257.16, 1122.94]
CWITE (Exponential)	591.72 [420.28, 795.09]	581.16 [365.49, 884.92]	**366.77 [303.75, 509.25]**	309.03 [254.19, 384.64]
CWITE (Linear)	**368.52 [338.49, 399.05]**	**232.73 [170.93, 318.22]**	536.77 [495.44, 574.41]	559.52 [486.90, 636.62]

the best baseline (7.01, 95% CI: [1.29, 12.93]) and the max-TTC ablation (9.70, 95% CI: [6.59, 14.30]) (**Appendix Table 10**). Similar trends appear in semi-synthetic data: both CWITE variants (linear: 368.52, 95% CI: [338.49, 399.05]; exponential: 591.72, 95% CI: [420.28, 795.09]) significantly outperform the best baseline (639.88, 95% CI: [543.41, 760.28]; top half of **Table 2**). These results demonstrate that CWITE remains robust and yields superior performance under violations of only Assumption 1, particularly when using a linear weighting function. Practical guidance for selecting g and sensitivity analyses for λ are reported in Appendix Section 5.

Table 3: Median MAE and 95% CIs for censored-like test individuals in the **synthetic-data setting with Assumption 2 violated** for earliest-TTE (Q1) and latest-TTE (Q5) quintiles. Assumption 1 holds.

Method	Censored-like Test Individuals with Earliest TTEs (in Q1)	Censored-like Test Individuals with Latest TTEs (in Q5)
IPCW	13.77 [1, 28]	9.16 [0, 27]
Standard	17.75 [1, 29]	**2.98 [0, 8]**
Standard Margin	12.90 [1, 25]	4.32 [0, 11]
CWITE	**3.01 [0, 8]**	5.40 [0, 13]

5.3. When Only Assumption 2 Is Violated

Assumption 2 requires each cluster to include at least one individual censored near their true TTE. This assumption is more likely to be violated when the error parameter associated with TTEs, η, is large and overall true TTEs in the dataset are large. To simulate this setting, we generated 1,000 synthetic datasets with larger η than in our original experiments and grouped test points by true TTE quintiles.

For individuals who would have been censored at training time in the latest-TTE quintile, CWITE's[1] MAE (5.4, 95% CI: [0, 13]) is worse than that of Standard (2.98, [0, 8]) and Standard Margin (4.32, [0, 11]), but better than IPCW (9.16, [0, 27]). This indicates that when only Assumption 2 is violated, CWITE loses its advantage relative to standard approaches. However, in the earliest-TTE quintile, CWITE (3.01, 95% CI: [0, 8]) signifi-

1. see Appendix Section 4 for details on the weighting function and other selected hyperparameters

cantly outperforms Standard (17.75, [1, 29]) and Standard Margin (12.9, [1, 25]) (**Table 3**), suggesting that its benefits persist when at least some censored individuals are observed near their true TTEs.

5.4. When Both Assumptions Are Violated

In semi-synthetic data with later TTEs (> 816 days), where both Assumption 1 and Assumption 2 are violated (verified in Appendix Section 5), no method clearly dominates. CWITE with exponential weighting achieves the lowest median MAE on censored-like test individuals (366.77, 95% CI: [303.75, 509.25]), while Standard performs best on uncensored-like test individuals (299.77, [245.30, 355.72]).

These results likely reflect the limited signal available to any method in this regime. IPCW depends entirely on reweighting uncensored training data, but in covariate regions with little uncensored support, there is insufficient data to reweight. Standard methods systematically overestimate TTEs for censored individuals; however, for individuals with later true TTEs, this overestimation is less harmful, as the true values are already large—which may explain why Standard remains competitive here. CWITE provides some supervision from censored data in regions where IPCW has none, and does not overestimate as aggressively as Standard, but when both assumptions are violated, neither mechanism provides reliable signal. Thus, while CWITE is competitive in this setting, its advantage is modest and no method clearly dominates (**Table 2**).

5.5. Real Data Case Study

As noted in the Experimental Setup Section, neither assumption can be verified directly here, though we mitigate violations of Assumption 1 through feature selection, and Assumption 2 remains plausible because interventions that censor spontaneous labor may be prompted by clinical features associated with labor timing, such as hypertension.

Under these conditions, CWITE[2] achieves a median MAE of 6.50 days (95% CI: [5.55, 7.40]) among uncensored test individuals who are similar to censored training individuals, outperforming the best baseline, which achieves 7.82 days (95% CI: [6.82, 8.82]). Among uncensored test individuals who are similar to uncensored training individuals, CWITE performs comparably to the best baseline: 6.50 days (95% CI: [5.54, 7.61]) versus 6.63 days (95% CI: [5.67, 7.69]) (**Table 4**). This result was robust to CWITE's clustering specification: varying k and replacing k-means with alternative clustering procedures produced only small changes in CWITE's MAE, which remained lower than that of all baselines (**Appendix Tables 14 and 15**).

To evaluate the potential relevance of these predictions for hospital resource planning, we also measure overprediction, defined as cases in which the predicted TTE exceeds the observed TTE and spontaneous labor therefore occurs earlier than predicted. In a scheduling context, such an error could result in an induction being scheduled after spontaneous labor has already occurred, rendering the planned intervention unnecessary. On the entire uncensored test set, compared to all evaluated methods, CWITE has the lowest overprediction rate (45.3% (95% CI: [44.5, 46.1]) vs. 49.4% (95% CI: [48.5, 50.1]) for IPCW, the best-performing baseline by MAE (**Appendix Table 18**)).

2. See Appendix Section 4 for details on the weighting function and selected hyperparameters.

Table 4: Median MAE (days) and 95% CI for test individuals in the **real-data setting** that are most similar to censored and uncensored training data after feature selection. In this setting, Assumption 1 likely does not hold, but Assumption 2 likely does.

Method	Censored-like Test Individuals	Uncensored-like Test Individuals
IPCW	7.82 [6.82, 8.82]	6.63 [5.67, 7.69]
Standard	12.49 [11.16, 13.90]	7.92 [6.74, 9.14]
Standard Margin	8.40 [7.40, 9.43]	7.24 [6.07, 8.56]
CWITE	**6.50** [**5.55, 7.40**]	**6.50** [**5.54, 7.61**]

Together with our earlier findings for settings in which only Assumption 1 is violated, these results suggest that CWITE can meaningfully improve predictive accuracy when Assumption 1 is unlikely to hold but Assumption 2 remains clinically plausible.

6. Discussion and Conclusion

Our work addresses two gaps in the survival analysis literature: settings where overlap between censored and uncensored training data is limited, as can occur when censoring is driven by deterministic interventions that systematically affect particular subgroups, and settings where the censoring mechanism changes across time or environments. Prior work has generally evaluated survival analysis methods on mixtures of individuals similar to uncensored and censored training data, which can obscure failures on the subgroups most affected by such shifts.

Empirically, CWITE consistently outperforms baselines on censored-like test individuals when Assumption 2 holds. When both Assumptions 1 and 2 hold in synthetic data, CWITE substantially reduces MAE on censored-like test individuals while maintaining comparable performance on uncensored-like test individuals (**Table 1**). This pattern persists when Assumption 1 is violated but Assumption 2 still holds, as in the semi-synthetic setting with early TTEs and, plausibly, in the real-data setting. In the real-data task of predicting time to spontaneous labor, CWITE reduces MAE (6.50 vs. 7.82 days; **Table 4**) and overprediction relative to IPCW (45.3% vs. 49.4%; **Appendix Table 18**). While this improvement is modest in absolute terms, it may still meaningfully support hospital resource planning, childcare coordination, and maternity leave scheduling.

A key limitation of CWITE arises when Assumption 2 is violated, as in the synthetic data setting with late TTEs, where clusters remain well-formed but all TTCs are far from TTEs. In this setting, CWITE performs slightly worse than the Standard approach. When both assumptions are violated, as in the semi-synthetic setting with later TTEs, no method clearly dominates: CWITE outperforms baselines at predicting TTEs for individuals similar to censored training data, but its advantage over baselines is modest. This highlights that CWITE is most effective when Assumption 2 holds.

Although Assumption 2 cannot be verified directly in real data because censored individuals' true TTEs are unobserved, it is plausible wherever the same observed trajectory influences both event timing and intervention timing: worsening clinical indicators may be associated with earlier spontaneous labor while also prompting induction or cesarean delivery, and analogous dynamics arise for transplantation before organ failure or prophylactic

intervention before disease onset. Because TTC is not determined by clinical risk alone—scheduling constraints, institutional workflows, and patient preferences also matter—some censored individuals with similar observed risk are observed close to when the event would otherwise have occurred. This is especially plausible for earlier events, where the possible TTC–TTE gap is smaller and accurate prediction is often most clinically important, as earlier labor onset is associated with greater neonatal risk (ACOG, 2021; Arnold, 2022; Knight et al., 2016; De Silva et al., 2017; Panelli et al., 2019). These findings therefore clarify CWITE's scope: it is most applicable when some censored observations plausibly occur near the events they replace, even if the learned clusters do not perfectly group individuals with similar true TTEs, and should not be expected to help when interventions consistently occur far before the event.

Our work has additional limitations. First, in the high-dimensional semi-synthetic setting, none of the clustering approaches evaluated—k-means, k-means with PCA, Gaussian mixture models (GMMs), or spectral clustering—satisfied Assumption 1. Although CWITE continued to perform well when Assumption 1 was violated, future work should investigate more robust clustering approaches that can better identify groups with similar true TTEs in high-dimensional clinical data.

Second, we considered only one model for the relationship between TTEs and TTCs (Assumption 3). Although our experiments vary the two aspects of that relationship most relevant to CWITE (within-cluster TTE similarity and the TTC–TTE gap), and Assumption 3 is not imposed in the real-data case study, they do not capture all possible relationships between event and censoring times. Future work should explore CWITE under a broader range of TTE/TTC data-generating processes.

In conclusion, we propose a survival analysis method designed for settings in which censoring is driven by deterministic interventions whose prevalence may shift over time. As intervention policies, clinical guidelines, and personal preferences continue to evolve (Keelan et al., 2021; Carmichael and Snowden, 2019; Press, 2012), methods that remain robust under censoring distribution shift will be increasingly important for reliable clinical decision-making.

7. Acknowledgements

M.K. received funding from the Graduate Fellowship for STEM Diversity. The views and conclusions in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of any funding sources. We thank Dr. Alex Peahl for helpful conversations about how a prediction of time to spontaneous labor would be used in practice, which informed our framing of the task and our choice of evaluation metrics. We also thank the anonymous reviewers for their valuable feedback.

References

ACOG. Medically indicated late-preterm and early-term deliveries, July 2021. URL https://www.acog.org/clinical/clinical-guidance/committee-opinion/articles/

2021/07/medically-indicated-late-preterm-and-early-term-deliveries. Committee Opinion No. 818.

Clarel Antoine and Bruce K. Young. Cesarean section one hundred years 1920–2020: the good, the bad and the ugly. *Journal of Perinatal Medicine*, 2020. doi: 10.1515/jpm-2020-0305. URL https://doi.org/10.1515/jpm-2020-0305.

Michael J. Arnold. Predicting and preventing preterm birth: Recommendations from acog. *American Family Physician*, 106(3):337–339, 2022. URL https://www.aafp.org/afp/practguide. Key recommendations include cessation of tobacco and alcohol use, increasing contraception access, and vaginal progesterone therapy for patients at risk.

Axios News. Severe flu season leads to increased hospitalizations in washington, d.c. *Axios*, 2025. URL https://www.axios.com/local/washington-dc/2025/03/03/flu-severe-season-kids-dc. Accessed: 2025-03-09.

Anders Bjorklund and Robert Moffitt. The estimation of wage gains and welfare gains in self-selection models. *The Review of Economics and Statistics*, 69(1):42–49, 1987. doi: 10.2307/1937899. URL https://doi.org/10.2307/1937899.

Emmanuel J. Candès, Lihua Lei, and Zhimei Ren. Conformalized survival analysis, 2023. URL https://arxiv.org/abs/2103.09763.

S.L. Carmichael and J.M. Snowden. The arrive trial: Interpretation from an epidemiologic perspective. *Journal of Midwifery & Women's Health*, 64(5):657–663, September 2019. doi: 10.1111/jmwh.12996. Epub 2019 Jul 2.

Victor Chernozhukov and Han Hong. Three-step censored quantile regression and extramarital affairs. *Journal of the American Statistical Association*, 97(459):872–882, 2002. doi: 10.1198/016214502388618663. URL https://doi.org/10.1198/0162145023886186.

Victor Chernozhukov, Iván Fernández-Val, and Amanda E. Kowalski. Quantile regression with censoring and endogeneity. *Journal of Econometrics*, 186(1):201–221, 2015. doi: 10.1016/j.jeconom.2014.06.017. URL https://doi.org/10.1016/j.jeconom.2014.06.017.

Dominiek Coates, Natasha Donnolley, Maralyn Foureur, and Amanda Henry. Inter-hospital and inter-disciplinary variation in planned birth practices and readiness for change: a survey study. *BMC Pregnancy and Childbirth*, 21(1):391, 2021. doi: 10.1186/s12884-021-03895-6.

Dane A. De Silva, Sarka Lisonkova, Peter von Dadelszen, Anne R. Synnes, Canadian Perinatal Network (CPN) Collaborative Group, and Laura A. Magee. Timing of delivery in a high-risk obstetric population: a clinical prediction model. *BMC Pregnancy and Childbirth*, 17:Article 202, 2017. doi: 10.1186/s12884-017-1396-3. Open access. Published: 29 June 2017.

Gaohong Dong, Lu Mao, Bo Huang, Margaret Gamalo-Siebers, Jiuzhou Wang, GuangLei Yu, and David C. Hoaglin. The inverse-probability-of-censoring weighting (ipcw) adjusted

win ratio statistic: An unbiased estimator in the presence of independent censoring. *Journal of Biopharmaceutical Statistics*, 30(5):882–899, September 2020. doi: 10.1080/10543406.2020.1757692. URL https://doi.org/10.1080/10543406.2020.1757692.

Bradley Efron and Robert J Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall/CRC, 1993.

Lisa Fields. Prenatal visit week 36, August 2024. URL https://www.webmd.com. Medically reviewed by Zilpah Sheikh, MD.

Lauren Gambino and Associated Press. Fda says it will limit access to covid-19 boosters for americans under 65, May 2025. URL https://www.theguardian.com/us-news/2025/may/20/fda-limits-covid-19-boosters. Accessed: 2025-06-03.

Junhao Gan and Yufei Tao. Dbscan revisited: Mis-claim, un-fixability, and approximation. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, SIGMOD '15, page 519–530, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450327589. doi: 10.1145/2723372.2737792. URL https://doi.org/10.1145/2723372.2737792.

Ali Hossein Foomani Gharari, Michael Cooper, Russell Greiner, and Rahul G Krishnan. Copula-based deep survival models for dependent censoring. In *Uncertainty in Artificial Intelligence*, pages 669–680. PMLR, 2023.

Miriam Gonzalez-Atienza, Dries Vanoost, Mathias Verbeke, and Davy Pissoort. Limitations of gaussian mixture models for symbol detection in harsh electromagnetic environments. *IEEE Letters on Electromagnetic Compatibility Practice and Applications*, 7(1):14–18, 2015. doi: 10.1109/LEMCPA.2024.3504712.

Nicholas Hartman, Sehee Kim, Kevin He, and John D Kalbfleisch. Pitfalls of the concordance index for survival outcomes. *Statistics in Medicine*, 42(13):2179–2190, 2023. doi: 10.1002/sim.9717.

James J. Heckman and Edward J. Vytlacil. Local instrumental variables and latent variable models for identifying and bounding treatment effects. *Proceedings of the National Academy of Sciences*, 96(8):4730–4734, 1999. doi: 10.1073/pnas.96.8.4730. URL https://doi.org/10.1073/pnas.96.8.4730.

Guido W. Imbens and Donald B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University, New York, 2015. ISBN 9780521885881.

Christopher H. Jackson. Flexsurv: A platform for parametric survival modeling in r. *Pharmaceutical Statistics*, 15(2):107–117, 2017. doi: 10.1002/pst.1747. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/pst.1947. Accessed: 2025-03-20.

Jared L Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. Deepsurv: Personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC Medical Research Methodology*, 18(1):24, 2018. doi: 10.1186/s12874-018-0482-1.

Stephen Keelan, Michael Flanagan, and Arnold D. K. Hill. Evolving trends in surgical management of breast cancer: An analysis of 30 years of practice changing papers. *Frontiers in Oncology*, 11:622621, 2021. doi: 10.3389/fonc.2021.622621. eCollection 2021.

William A. Knaus, Frank E. Harrell, Joanne Lynn, and et al. The support prognostic model: Objective estimates of survival for seriously ill hospitalized adults. *Annals of Internal Medicine*, 122(3):191–203, 1995. doi: 10.7326/0003-4819-122-3-199502010-00007.

Hannah E. Knight, Jan H. van der Meulen, Ipek Gurol-Urganci, Gordon C. Smith, Amit Kiran, Steve Thornton, David Richmond, Alan Cameron, and David A. Cromwell. Birth "out-of-hours": An evaluation of obstetric practice and outcome according to the presence of senior obstetricians on the labour ward. *PLOS Medicine*, 13(4):e1002000, 2016. doi: 10.1371/journal.pmed.1002000. URL https://doi.org/10.1371/journal.pmed.1002000.

Amanda E. Kowalski. Censored quantile instrumental variable estimates of the price elasticity of expenditure on medical care. *Journal of Business & Economic Statistics*, 34 (1):107–117, 2016. doi: 10.1080/07350015.2015.1004072. URL http://dx.doi.org/10.1080/07350015.2015.1004072.

Amanda E. Kowalski. Reconciling seemingly contradictory results from the oregon health insurance experiment and the massachusetts health reform. *The Review of Economics and Statistics*, 105(3):646–664, 2023. doi: 10.1162/rest_a_01069. URL https://doi.org/10.1162/rest_a_01069.

Changhee Lee, William Zame, Jinsung Yoon, and Mihaela van der Schaar. Deephit: A deep learning approach to survival analysis with competing risks. *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, 32(1), 2018. doi: 10.1609/aaai.v32i1.11842. URL https://doi.org/10.1609/aaai.v32i1.11842.

Changhee Lee, Jinsung Yoon, and Mihaela van der Schaar. Dynamic-deephit: A deep learning approach for dynamic survival analysis with competing risks based on longitudinal data. *IEEE Transactions on Biomedical Engineering*, 67(1):122–133, 2020. doi: 10.1109/TBME.2019.2909027. Epub 2019 Apr 3.

Fan Li, Laine E. Thomas, and Fan Li. Addressing extreme propensity scores via the overlap weights. *American Journal of Epidemiology*, 188(1):250–257, 2018. doi: 10.1093/aje/kwy201. URL https://academic.oup.com/aje/article/188/1/250/5090958.

Emilio Tanowe Maddalena, Paul Scharnhorst, and Colin N. Jones. Deterministic error bounds for kernel-based learning techniques under bounded noise. *Automatica*, 134: 109896, 2021. ISSN 0005-1098. doi: https://doi.org/10.1016/j.automatica.2021.109896. URL https://www.sciencedirect.com/science/article/pii/S0005109821004192.

Mayo Clinic Staff. Puberty blockers for transgender and gender-diverse youth, June 2023. URL https://www.mayoclinic.org/diseases-conditions/gender-dysphoria/in-depth/pubertal-blockers/art-20459075. Accessed: 2025-06-03.

Kelly Metcalfe, Andrea Eisen, Leigha Senter, Susan Armel, Louise Bordeleau, Wendy S. Meschino, Tuya Pal, Henry T. Lynch, Nadine M. Tung, Ava Kwong, Peter Ainsworth,

Beth Karlan, Pal Moller, Charis Eng, Jeffrey N. Weitzel, Ping Sun, Jan Lubinski, and Steven A. Narod. International trends in the uptake of cancer risk reduction strategies in women with a brca1 or brca2 mutation. *British Journal of Cancer*, 121:15–21, 2019. doi: 10.1038/s41416-019-0451-1.

Chirag Nagpal, Xinyu Li, and Artur Dubrawski. Deep survival machines: A fully parametric survival regression model for censored data. In *Proceedings of the Machine Learning for Healthcare Conference*, volume 149 of *Proceedings of Machine Learning Research*, page N/A. PMLR, 2021. URL https://proceedings.mlr.press/v149/nagpal21a.html.

Arash Pakbin, Xiaochen Wang, Bobak J. Mortazavi, and Donald K. K. Lee. Boxhed 2.0: Scalable boosting of functional data in survival analysis. *CoRR*, abs/2103.12591, 2021. URL https://arxiv.org/abs/2103.12591.

Danielle M. Panelli, Julian N. Robinson, Anjali J. Kaimal, Kathryn L. Terry, Jiaxi Yang, Mark A. Clapp, and Sarah E. Little. Using cervical dilation to predict labor onset: A tool for elective labor induction counseling. *American Journal of Perinatology*, 36(14): 1485–1491, 2019. doi: 10.1055/s-0039-1677866. Original Article. Funding: This study was partially supported by an Expanding the Boundaries grant through Brigham and Women's Hospital.

Dale A. Patterson, Marguerite Winslow, and Coral D. Matus. Spontaneous vaginal delivery. *American Family Physician*, 78(3):336–341, August 2008. URL https://www.aafp.org. A more recent article on the management of spontaneous vaginal delivery is available.

L. Pereira and A. B. Caughey. Predicting preterm birth: where do the major challenges lie? *Journal of Perinatology*, 32:481–482, 2012. doi: 10.1038/jp.2012.44. 500 Accesses, 1 Citation, 2 Altmetric.

James L Powell. Least absolute deviations estimation for the censored regression model. *Journal of Econometrics*, 25(3):303–325, 1984. ISSN 0304-4076. doi: https://doi.org/10.1016/0304-4076(84)90004-6. URL https://www.sciencedirect.com/science/article/pii/0304407684900046.

Associated Press. Hospitals crack down on induced labors, 2012. URL https://www.nbcnews.com/health/health-news/hospitals-crack-down-induced-labors-flna1c9453799. Accessed: 2024-12-13.

Shi-ang Qi, Neeraj Kumar, Mahtab Farrokh, Weijie Sun, Li-Hao Kuan, Rajesh Ranganath, Ricardo Henao, and Russell Greiner. An effective meaningful way to evaluate survival models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*. PMLR, 2023.

James M Robins and Andrea Rotnitzky. Recovery of information and adjustment for dependent censoring using surrogate markers. In *AIDS epidemiology: methodological issues*, pages 297–331. Springer, 1992.

Karen R. Rosenberg and Wenda R. Trevathan. Evolutionary perspectives on cesarean section. *Evolution, Medicine, and Public Health*, 2018(1):67–81, 2018. doi: 10.1093/emph/eoy006. URL https://doi.org/10.1093/emph/eoy006.

Shengpu Tang, Parmida Davarmanesh, Yanmeng Song, Danai Koutra, Michael W. Sjoding, and Jenna Wiens. Democratizing ehr analyses with fiddle: a flexible data-driven preprocessing pipeline for structured clinical data. *JAMIA Open*, 3(1):80–91, 2020.

The Leapfrog Group. State of maternity care in u.s. hospitals. Technical report, The Leapfrog Group, March 2025. URL https://www.leapfroggroup.org/sites/default/files/Files/MaternityCare-report-2025-FINAL_0.pdf. Accessed: 2025-06-03.

Laine E. Thomas, Fan Li, and Michael J. Pencina. Overlap weighting: A propensity score method that mimics attributes of a randomized clinical trial. *JAMA*, 323(23):2417–2418, 2020. doi: 10.1001/jama.2020.7819. URL https://jamanetwork.com/journals/jama/article-abstract/2765686.

Donna Tjandra, Yifei He, and Jenna Wiens. A hierarchical approach to multi-event survival analysis. In *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21)*, Virtual Conference, 2021. AAAI Press, Association for the Advancement of Artificial Intelligence.

VISTA Staffing Solutions. Doctor shortages during the holidays impact hospital revenue: How locum tenens can help, September 2024. URL https://www.vistastaff.com/workforce-optimization/doctor-shortages-during-the-holidays-impacts-hospital-revenue/. Accessed: 2025-06-03.

Zifeng Wang and Jimeng Sun. Survtrace: transformers for survival analysis with competing events. In *Proceedings of the 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, BCB '22, page 1–9. ACM, August 2022. doi: 10.1145/3535508.3545521. URL http://dx.doi.org/10.1145/3535508.3545521.

Weijia Zhang, Chun Kai Ling, and Xuanhui Zhang. Deep copula-based survival analysis for dependent censoring with identifiability guarantees. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 20613–20621, 2024.

Yunji Zhou, Roland A Matsouaka, and Laine Thomas. Propensity score weighting under limited overlap and model misspecification. *Statistical Methods in Medical Research*, 29(12):3691–3705, 2020. doi: 10.1177/0962280220940334. URL https://doi.org/10.1177/0962280220940334.

Appendix Section 1: Problem Setup

In alignment with past work, we make the following assumptions

Assumption 4 (Right Censoring) *We assume that if an individual is censored* ($e^{(i)} = 0$)*, their TTC precedes their true TTE:* $t_C^{(i)} < t_E^{(i)}$.

This condition holds when censoring occurs due to preventive interventions that necessarily predate a medical event.

Furthermore, like previous work, we also assume identifiability of TTEs and e from x alone (Lee et al., 2018; Robins and Rotnitzky, 1992; Dong et al., 2020; Wang and Sun, 2022).

Assumption 5 (Identifiability from Observed Covariates) *We assume that the observed covariates* $\mathbf{x}^{(i)}$ *are sufficient to model both the TTE and censoring mechanism. Formally, letting* $\mathbf{u}^{(i)}$ *denote unobserved variables, we assume*

$$t_E^{(i)} \perp\!\!\!\perp \mathbf{u}^{(i)} \mid \mathbf{x}^{(i)}, \quad e^{(i)} \perp\!\!\!\perp \mathbf{u}^{(i)} \mid \mathbf{x}^{(i)}.$$

In the absence of this assumption, methods typically rely on strong parametric models that specify the relationship between $\mathbf{x}$, e, and t_E (Gharari et al., 2023; Zhang et al., 2024). In contrast, we focus on data-driven learning approaches that aim to flexibly model censoring and the TTE distributions based on observed covariates. To enable this flexibility while avoiding parametric assumptions, we adopt Assumption 5.

Appendix Section 2: Related Work

In this work, we focus on developing machine learning–based survival analysis methods that produce point estimates of expected TTEs. Recent work has increasingly shifted toward alternative objectives, such as constructing prediction intervals using conformal prediction or estimating conditional quantiles. These approaches are designed for uncertainty quantification rather than point estimation, and often rely on assumptions that are not suitable for our setting.

Quantile regression assumes that individuals with low censoring probabilities within each quantile are representative of the quantile in the test set (Chernozhukov and Hong, 2002; Chernozhukov et al., 2015; Kowalski, 2016). Conformal prediction methods require a calibration set that is representative of the test set to ensure valid coverage guarantees (Candès et al., 2023). Both assumptions break down under censoring distribution shift, where the relationship between censoring and covariates changes across time or environments—such that individuals censored in the calibration set may not be censored in the test set, and vice versa. To our knowledge, existing conformal survival methods do not account for such shifts.

A separate line of work addresses unobserved confounding by introducing latent variables or structural assumptions about the data-generating process (Gharari et al., 2023; Zhang et al., 2024). These methods typically rely on strong parametric assumptions about the relationships between covariates, latent variables, and TTEs. In contrast, we focus on settings without unobserved confounding, which simplifies the problem and avoids the risk

of model misspecification. Moreover, these latent-variable approaches are likely to fail when the timing of clinical events deviates from assumed generative structures—a common occurrence in healthcare, where event timing is driven by complex, often poorly understood biological processes (Jackson, 2017).

As discussed in the Problem Formalization & Background Section, standard ML-based survival methods relevant to our setting assume uninformative censoring and are trained to (1) match observed event times for uncensored individuals, and (2) predict TTEs greater than censoring times for censored individuals (Lee et al., 2018; Wang and Sun, 2022). These models tend to overestimate TTEs for censored individuals in regions with few uncensored examples, since overestimation goes unpenalized (see Figure 1). IPCW attempts to address this by reweighting uncensored individuals to approximate the full population (Robins and Rotnitzky, 1992; Dong et al., 2020), but its effectiveness depends on sufficient covariate overlap. In settings with limited overlap—a common scenario in clinical datasets—IPCW becomes unstable, even with regularization strategies such as weight clipping (Imbens and Rubin, 2015; Li et al., 2018; Zhou et al., 2020; Thomas et al., 2020).

In summary, existing approaches either lack meaningful supervision for censored individuals (standard models) or rely on strong assumptions that break down under censoring shift and limited overlap (quantile/conformal methods, IPCW, and latent-variable models). CWITE addresses these limitations by enabling specific supervision from censored training data without requiring overlap or strong assumptions about the data-generating process.

Appendix Section 3: Methods

Pseudocode for CWITE Algorithm 1 shows the pseudocode for CWITE.

Algorithm 1 Pseudocode for CWITE

Inputs: x^i, e^i, t^i_{obs}, n_c, $g(\cdot)$
Cluster the covariates in x into n_c clusters.
Initialize vector `loss`
for each i **do**
  `loss`[i] $= \mathcal{L}\left(\hat{t}_e^{(i)}, t_{obs}^{(i)}\right)$, the NLL loss between predicted TTE and observed time
  if $e^{(i)} = 0$ **then**
    $t^{(i)}_{\max} = \max\{t^{(j)}_{obs} \mid j \text{ in same cluster as } i, e^{(j)} = 0\}$
    `loss`[i] $= \frac{g\left(t_{obs}^{(i)}\right)}{g(t^{(i)}_{max})}$`loss`[i]
  else
    p_{uncens}[i] $= P(e^{(i)} \mid x^{(i)})$
    `loss`[i] $= \frac{1}{p_{\text{uncens}}[\text{i}]}$`loss`[i]
  end if
end for
Return `loss`

Motivation for DGP in Assumption 3. We illustrate the assumed data-generating process (DGP) for Assumption 3 in Figure 2. We assume that both $t_E^{(i)}$ and $t_C^{(i)}$ are generated

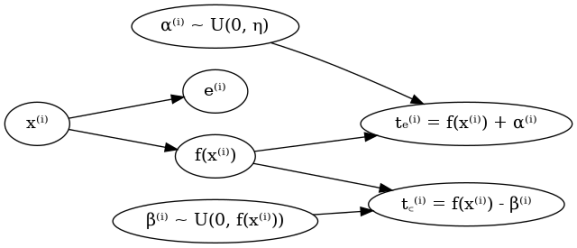

Figure 2: Illustration of the assumed data-generating process (DGP) in Assumption 3

from a shared latent function $f(\mathbf{x}^{(i)})$, perturbed by independent uniform noise:

$$t_E^{(i)} = f(\mathbf{x}^{(i)}) + \alpha^{(i)}, \quad t_C^{(i)} = f(\mathbf{x}^{(i)}) - \beta^{(i)},$$

where $\alpha^{(i)} \sim \mathcal{U}(0, \eta)$ and $\beta^{(i)} \sim \mathcal{U}(0, f(\mathbf{x}^{(i)}))$, ensuring that $t_C^{(i)} < t_E^{(i)}$ and thus right censoring occurs.

An arguably more intuitive DGP would draw $t_C^{(i)} \sim \mathcal{U}(0, t_E^{(i)})$, guaranteeing that censoring always precedes the event. However, this introduces unobserved confounding between t_E and t_C, violating the conditional independence assumption needed for principled estimation. Our formulation avoids this issue by ensuring that $t_E^{(i)} \perp t_C^{(i)} \mid \mathbf{x}^{(i)}$.

Appendix Section 4: Experimental Setup

Synthetic Data

First, we generate 1000 datasets, each defined by a unique functional relationship between the covariate and f, specified by weights $w_{s,1}, w_{c,1}, w_{s,2}, w_{c,2}$ sampled once per dataset. These weights remain fixed throughout the dataset and introduce variation across datasets. Overlap is additionally controlled by parameters $\epsilon_o \in \{0.1, 0.5, 1.0\}$ and $\tau \in \{0.5, 0.75, 0.9\}$.

Each dataset contains $n = 10{,}000$ samples, equally split into training, validation, and test sets. Each individual is assigned a single covariate sampled as $x^{(i)} \sim \mathcal{U}([-3, 3])$.

Generating f. The function $f(\mathbf{x}^{(i)})$ we use to generate TTEs and TTCs is

$$\left\lfloor \frac{w_{s,1} \sin(w_{s,2} \cdot x^{(i)}) + w_{c,1} \cos(w_{c,2} \cdot x^{(i)}) - z_{\min}}{z_{\max} - z_{\min}} \cdot 30 \right\rfloor,$$

where $w_{s,1}, w_{c,1} \sim \mathcal{U}([1, 20])$, $w_{s,2}, w_{c,2} \in \{1, 2\}$, and $z_{\min} = -w_{s,1} - w_{c,1}$, $z_{\max} = w_{s,1} + w_{c,1}$. $w_{s,1}$ and $w_{c,1}$ are chosen to ensure that f varies meaningfully with the covariate. $z_{\min}$ and $z_{\max}$ normalize the output to a consistent range, facilitating stable learning.

Censoring Mechanism e. We define a stochastic censoring mechanism applied only to training and validation data:

$$e^{(i)} = \mathbb{1}\left(\frac{1}{1 + e^{-x^{(i)} + \eta^{(i)}}} > \tau \right)$$

$$\eta^{(i)} \sim \mathcal{U}([-\epsilon_o, \epsilon_o])$$

The parameters $\epsilon_o \in \{0.1, 0.5, 1.0\}$ control the level of noise in the event indicator, and $\tau \in \{0.5, 0.75, 0.9\}$ sets the censoring threshold. Larger ϵ_o increases overlap; higher τ results in more censoring and less overlap.

Violating Assumption 1. To isolate the effect of poor cluster quality, we reuse the first set of 1,000 baseline synthetic datasets ($\eta \in \{1, 2, 5\}$) but randomly reassign 25% of individuals' cluster memberships after k-means clustering. For each reassigned individual, we sample a new cluster label uniformly from the remaining clusters. This preserves the underlying data-generating process — and thus Assumption 2 — while directly degrading within-cluster TTE homogeneity. All DGP parameters (η, ϵ_o, τ) remain unchanged.

Violating Assumption 2. The second set of 1,000 synthetic datasets uses the same data-generating process as the baseline but with larger TTE noise: $\eta \in \{5, 10\}$ (compared to $\eta \in \{1, 2, 5\}$ in the baseline). Larger η increases the spread of $\Delta^{(i)} = t_E^{(i)} - t_C^{(i)}$, making it less likely that any censored individual is censored within ϵ of their true TTE. Clustering is performed normally (no perturbation), so Assumption 1 holds. We evaluate results stratified by true TTE quintile, since as shown in the Theoretical Analysis, Assumption 2 is more likely to hold for individuals with earlier TTEs.

Semi-Synthetic Data

To evaluate CWITE under more realistic conditions, we generate semi-synthetic data using real covariates and observed TTEs. Qi et al. (2023) include multiple real-world survival analysis datasets that can be used to generate semi-synthetic data: GBM, METABRIC, SUPPORT, and two based on MIMIC-IV. In this work, we generate semi-synthetic data using the SUPPORT dataset because METABRIC and GBM contain $\leq$ 1,000 uncensored individuals, leading to high variance in model estimates. Meanwhile, the MIMIC-IV-based datasets exhibit heavily skewed event-time distributions, making TTE prediction trivial for most approaches. The SUPPORT dataset (Knaus et al., 1995) contains clinical covariates $\mathbf{x}^{(i)} \in \mathbb{R}^{31}$ and time-to-death labels ranging from 3 to 1944 days. We define the semi-synthetic censoring split using clinically meaningful severity markers from SUPPORT: individuals are censored in training and validation if coma hours > 30, bilirubin > 1.0, creatinine > 1.2, mean blood pressure < 80, or albumin < 3.4. This covariate-driven split is appropriate for the semi-synthetic setting because it creates limited uncensored support using observed clinical risk factors related to severity, while preserving a fully observed test set for evaluating censoring distribution shift.

Generating f. Using all uncensored training and validation data, f is a neural network fit to learn relationship between observed covariates and original TTEs. The TTEs that we consider ground truth have small uniform noise added to f.

Censoring Mechanism e. We apply censoring only to the training and validation sets. Our mechanism e depends on clinical markers of illness severity—coma hours (`scoma`), bilirubin (`bili`), creatinine (`crea`), mean blood pressure (`meanbp`), and albumin (`alb`)—used in SOFA/APACHE II scores. An individual is censored if *any* adverse threshold is exceeded, and is therefore uncensored only if all of the corresponding non-adverse inequalities hold:

$$e^{(i)} = \mathbf{1}\left\{\texttt{scoma}_i \leq 30 \ \wedge \ \texttt{bili}_i \leq 1.0 \ \wedge \ \texttt{crea}_i \leq 1.2 \ \wedge \ \texttt{meanbp}_i \geq 80 \ \wedge \ \texttt{alb}_i \geq 3.4\right\}$$

Feature	Range	Overall (N = 45632)	Spontaneous Labor (N = 21645)
Age	12.0 - 19.0	502, (1.10 %)	225, (1.04 %)
Age	19.0 - 24.0	4233, (9.28 %)	2092, (9.67 %)
Age	24.0 - 34.0	27519, (60.31 %)	13045, (60.27 %)
Age	34.0 - 44.0	13160, (28.84 %)	6178, (28.54 %)
Age	44.0 - 55.7	218, (0.48 %)	105, (0.49 %)
Ethnicity	Hispanic or Latino	2675, (5.86 %)	1269, (5.86 %)
Ethnicity	Non-Hispanic or Latino	41957, (91.95 %)	19923, (92.04 %)
Ethnicity	Patient Refused	243, (0.53 %)	107, (0.49 %)
Ethnicity	Unknown	585, (1.28 %)	256, (1.18 %)
# Fetuses in Pregnancy	> 1	663, (1.45 %)	298, (1.38 %)
# Fetuses in Pregnancy	1	44969, (98.55 %)	21347, (98.62 %)
# Total Pregnancies	> 1	32359, (70.91 %)	15271, (70.55 %)
# Total Pregnancies	1	13273, (29.09 %)	6374, (29.45 %)
Race	African American	5726, (12.55 %)	2748, (12.70 %)
Race	American Indian	167, (0.37 %)	83, (0.38 %)
Race	Asian	3731, (8.18 %)	1786, (8.25 %)
Race	Caucasian	32768, (71.81 %)	15522, (71.71 %)
Race	Other	2316, (5.08 %)	1088, (5.03 %)
Race	Patient Refused	255, (0.56 %)	115, (0.53 %)
Race	Unknown	347, (0.76 %)	153, (0.71 %)

Table 5: Characteristics of the data used to train, validate, and test approaches to predicting time to spontaneous labor

Real Data

Study Cohort. Our study cohort includes pregnant individuals with encounters at Michigan Medicine, the University of Michigan's academic medical center, between January 1, 2014 and December 31, 2024. Individuals were included if they remained pregnant through at least the 36th week of gestation; those who delivered prior to 36 weeks were excluded (Figure 3). We define the prediction time as the start of the 36th gestational week, a clinically meaningful point at which healthcare providers begin offering highly personalized care plans (Fields, 2024). Although earlier prediction would be ideal, forecasting premature labor has been shown to be substantially more difficult than predicting labor later in pregnancy (Pereira and Caughey, 2012). After applying inclusion and exclusion criteria, the final cohort consisted of 45,632 individuals. Baseline characteristics of the cohort used for model development and evaluation are summarized in Table 5.

Outcome Definition. An individual is labeled as having experienced spontaneous labor ($e = 1$) if a "labor" status is recorded in their electronic health record prior to delivery, not preceded by "induction." The TTE is the timestamp of the first "labor" record. All others

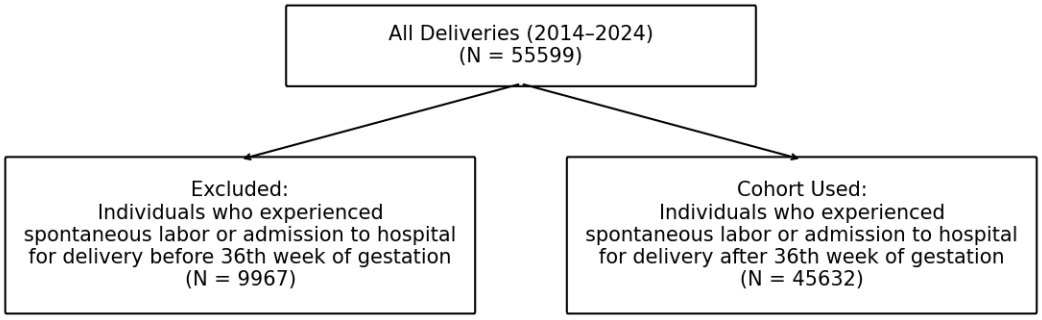

Figure 3: Study cohort inclusion criteria for predicting time to spontaneous labor

are considered censored ($e = 0$), with their observed time set to the admission timestamp prior to delivery.

Feature Extraction and Preprocessing. The data used to train and evaluate our model include individuals' vitals and diagnoses recorded prior to the start of the 36th gestational week, along with several time-invariant features: demographics, estimated due date, date of last menstrual period, number of previous pregnancies, and number of fetuses in the current pregnancy.

Time-invariant features were processed using FIDDLE with its default preprocessing settings (Tang et al., 2020). Time-varying features were handled as follows. For binary variables (*e.g.*, diagnoses), we extracted the most recent observation recorded during each of the following periods: pre-pregnancy, first trimester, second trimester, and third trimester (prior to 36 weeks). For continuous variables (*e.g.*, vitals), we extracted the most recent value from each period as well, and additionally computed summary statistics—minimum, maximum, mean, standard deviation, and temporal slope—across all available measurements prior to 36 weeks.

Missing covariate values were imputed using a dummy fill-in approach, supplemented by the addition of indicator variables to flag missingness for each feature.

Implementation Details

In line with previous work, all methods evaluated in our work use MLP-based feature extractors (Lee et al., 2018; Katzman et al., 2018; Nagpal et al., 2021; Tjandra et al., 2021). Existing survival analysis approaches often incorporate ranking- or calibration-based objectives alongside TTE-prediction losses (Lee et al., 2018). However, recent work has emphasized that the error between predicted and true TTEs is the most clinically meaningful evaluation metric in survival analysis (Qi et al., 2023). Motivated by this, we focus exclusively on integrating censored individuals into loss functions that directly optimize prediction accuracy. Our comparisons therefore isolate the effect of the objective function while holding the model architecture fixed. CWITE is architecture-agnostic and can

Table 6: Hyperparameter search spaces and fixed settings. All neural models were optimized with Adam. For IPCW, the propensity model used logistic regression with C selected by validation AUROC.

Experiment	Methods	Shared architecture and training settings	CWITE-specific settings
Synthetic	IPCW, Standard, Standard Margin, CWITE	3 hidden layers; hidden size $\{16, 32, 64\}$; batch size $\{128, 256, 512, 1024\}$; learning rate $\{10^{-2}, 10^{-3}\}$; weight decay $\{10^{-4}, 10^{-5}, 10^{-6}\}$; 500 epochs with patience 15; five random draws for baselines; IPCW $C \in \{1, 10^{-2}, 10^{-4}, 10^{-6}\}$	k selected by the elbow method; $\lambda \in \{0.01, 0.05, 0.1, 0.5, 1\}$; $g \in \{\text{linear}, \text{exponential}, \text{polynomial}\}$; ten random draws
Semi-synthetic	IPCW, Standard, Standard Margin, CWITE	Hidden layers $\{1, 2, 3\}$; hidden size $\{16, 32, 64\}$; batch/learning-rate pairs: $64/\{10^{-3}, 5\times10^{-4}\}$, $128/\{5\times10^{-4}, 10^{-4}\}$, $256/\{10^{-4}\}$; weight decay $\{10^{-4}, 10^{-5}, 10^{-6}\}$; 500 epochs with patience 15; 25 random draws; seed 42; IPCW $C \in \{1, 10^{-2}, 10^{-4}, 10^{-6}\}$	Main CWITE configuration selected $k = 95$, $g =$ linear, and $\lambda = 0.075$; additional sensitivity analyses varied g and λ (**Table 17**)
Real data	IPCW, Standard, Standard Margin, CWITE	3 hidden layers; hidden size 64; batch size 128; learning rate 10^{-4}; weight decay 10^{-4}; 200 epochs with patience 15; seed 42; feature transformations selected from standardized all features or SelectKBest top $\{50, 100, 200\}$ features	$k \in \{5, 10, 20\}$; $g =$ linear; $\lambda \in \{0.05, 0.1, 0.25, 0.5, 0.7, 0.8, 0.9, 1.0\}$; normalized censored-loss scaling; clustering sensitivity used the alternatives reported in Appendix Section 5

Table 7: Selected hyperparameters for the reported semi-synthetic SUPPORT severity-marker split.

Method	Layers	Hidden	Batch	LR	WD	C	k	g	λ
IPCW	1	32	128	10^{-4}	10^{-6}	1.0	–	–	–
Standard	1	64	64	10^{-3}	10^{-5}	–	–	–	–
Standard Margin	3	32	256	10^{-4}	10^{-6}	–	–	–	–
CWITE	1	16	128	5×10^{-4}	10^{-6}	1.0	95	linear	0.075

be combined with deeper or attention-based representations, as well as with ranking- and calibration-based objectives.

Table 6 reports the hyperparameter search spaces and fixed settings used in each experiment. Synthetic experiments used independently generated replicates, so final selected architectures varied by replicate; the full replicate-level selected configurations are provided in the released code rather than printed in the manuscript. For the semi-synthetic experiments, we report the selected hyperparameters for the reported SUPPORT severity-marker split (**Table 7**). The real-data experiments used a fixed architecture for all methods and selected among feature transformations and, for CWITE, k and λ using uncensored validation MAE; exact selected real-data settings are shown in Table 8.

In our synthetic data experiments – when both Assumption 1 and 2 hold, when only Assumption 1 holds, and when only Assumption 2 holds – we swept over overall weight magnitudes and transformation functions. In the semi-synthetic setting, we evaluated CWITE

Table 8: Selected real-data architecture and hyperparameter settings. All methods used the same fixed neural architecture; validation selected among standardized all features and SelectKBest top 50, 100, or 200 features.

Method	Layers	Hidden	Batch	LR	WD	Feature settings	k	g, λ
IPCW	3	64	128	10^{-4}	10^{-4}	SelectKBest top 100	–	–
Standard	3	64	128	10^{-4}	10^{-4}	SelectKBest top 200	–	–
Standard Margin	3	64	128	10^{-4}	10^{-4}	SelectKBest top 200	–	–
CWITE	3	64	128	10^{-4}	10^{-4}	SelectKBest top 200	5	linear, 1.0

using both linear and exponential weighting, each with small and large λ; linear weighting outperformed exponential weighting on MAE (**Table 2**) and, under linear weighting, performance was largely insensitive to λ (**Table 17**). Based on this, we used linear weighting in our real-data experiments. In practice, the weighting function g should reflect the expected reliability of the clusters and cluster maximum TTCs: exponential weighting is preferable in lower-dimensional settings where these quantities are likely informative, while linear weighting is safer in high-dimensional settings with noisier clusters. The overall weight λ should reflect the plausibility of close-to-event censoring and the prediction priority: larger λ is appropriate when TTCs are likely close to true TTEs or early-TTE accuracy is the priority, whereas smaller λ is preferable when TTCs may be far from true TTEs.

For the synthetic and semi-synthetic data experiments, we performed random splits of the data into training, validation, and test sets. For the real-data task, we aimed to create a setting where the probability of censoring given covariates was higher in training than in test, as described in the Problem Formalization & Background Section. Observing a decline in spontaneous labor rates over time, we split the data temporally—assigning individuals from later years (2021–2024) to the training set and those from earlier years (2014–2020) to the test set. The validation set is drawn as a random hold-out from the training-year (2021–2024) individuals, so its censoring distribution matches the training set rather than the test set; we use it to select hyperparameters by optimizing MAE on its uncensored individuals. Although this design reverses the typical retrospective setup, it simulates a realistic deployment setting where the probability of censoring may decrease over time due to evolving clinical practices.

In CWITE, we implement clustering using k-means due to its simplicity and computational efficiency. In Appendix Section 5, we compare k-means to more complex clustering methods—including Gaussian Mixture Models (GMMs), and spectral clustering—by evaluating their ability to satisfy Assumption 1 (Gonzalez-Atienza et al., 2015; Gan and Tao, 2015) (**Table** 13). We also evaluate whether real-data results are sensitive to the number of k-means clusters or to replacing k-means with alternative clustering procedures.

All experiments were conducted on a dedicated server equipped with eight NVIDIA GeForce RTX 2080 Ti GPUs (each with 12.9 GB of RAM) and a single AMD EPYC 7702 processor with 64 physical cores and 128 threads, based on the Zen 2 (Rome) architecture.

Code Availability. Code for the synthetic, semi-synthetic, and real-data analyses is available at https://github.com/MLD3/CWITE. The repository includes scripts for reproducing the synthetic and semi-synthetic experiments, scripts documenting the real-data

analysis pipeline, and helper scripts for recovering selected hyperparameters from experiment outputs. The protected clinical data used in the real-data experiments cannot be redistributed.

Appendix Section 5: Results

Verifying Assumptions In our synthetic and semi-synthetic settings, right censoring and identifiability of t_E, t_C, and censoring status from observed covariates hold by design.

Assumption 1 states that the clustering procedure groups individuals such that within each cluster k, true TTEs are relatively homogeneous. In contrast, variation in TTCs among censored individuals may reflect additional censoring noise (e.g., due to administrative factors), inflating their variance beyond that of the underlying TTEs. Formally, for each cluster k, we expect

$$\mathrm{Var}_{i \in C_k^c}(t_C^{(i)}) - \mathrm{Var}_{i \in C_k^c}(t_E^{(i)}) > 0.$$

Assumption 2 states that each cluster contains at least one censored individual observed sufficiently close to their true TTE. Formally,

$$P\left(\min_{i \in C_k^c} \Delta^{(i)} \leq \epsilon\right) \gg 0,$$

where $\Delta^{(i)} = t_E^{(i)} - t_C^{(i)}$.

To empirically assess these assumptions, we compute two worst-case cluster-level diagnostics over censored individuals: (1) the minimum variance difference,

$$\min_k \left(\mathrm{Var}_{i \in C_k^c}(t_C^{(i)}) - \mathrm{Var}_{i \in C_k^c}(t_E^{(i)})\right),$$

and (2) the maximum minimum TTC–TTE gap within a cluster,

$$\max_k \left(\min_{i \in C_k^c} \left|t_C^{(i)} - t_E^{(i)}\right|\right).$$

These diagnostics should be interpreted as matters of degree rather than binary indicators. The first quantity is strictly positive only if Assumption 1 holds in every cluster; values well above zero provide strong support for the assumption, values near zero indicate the assumption is approximately satisfied, and increasingly negative values indicate progressively stronger violations. Similarly, the second quantity is small only if every cluster contains at least one censored individual close to their true TTE; values near zero strongly support Assumption 2, while larger values reflect progressively weaker support. Because both quantities are continuous, we interpret them in relation to one another and to baseline values rather than as strict pass/fail criteria.

In the synthetic setting, where we evaluate many datasets, we compute these worst-case metrics for each dataset and report their mean across datasets. In the semi-synthetic setting, which consists of a single dataset, we fit clusters on the training covariates, assign validation points to those clusters, and report the corresponding worst-case metrics separately for the early- and late-TTE subsets.

Table 9: Verification of clustering assumptions. For the synthetic settings, we report the mean across datasets of two worst-case cluster-level diagnostics. For the semi-synthetic setting, which consists of a single dataset, we report the corresponding worst-case metric computed within each subset. Positive values in column 3 support Assumption 1; small values in column 4 support Assumption 2.

Setting	Data Subset	$\min_k (\mathrm{Var}(t_C) - \mathrm{Var}(t_E))$	$\max_k \left(\min_i \lvert t_C^{(i)} - t_E^{(i)} \rvert\right)$
Synthetic (baseline, $\eta \in \{1, 2, 5\}$)	All	0.51	2.65
Synthetic (clusters perturbed 25%)	All	−39.42	1.62
Synthetic (large $\eta \in \{5, 10\}$)	All	−3.93	4.42
Semi-Synthetic	$t_E < 816$ days	−25.34	0.00
	$t_E \geq 816$ days	−30.44	93.97

In the baseline synthetic setting ($\eta \in \{1, 2, 5\}$), both assumptions hold on average across datasets: the mean minimum variance difference is 0.51 and the mean maximum minimum gap is 2.65. When cluster assignments are perturbed by 25% to violate Assumption 1, the mean minimum variance difference drops to -39.42, confirming that within-cluster TTE homogeneity is degraded, while the mean maximum minimum gap remains low at 1.62, indicating that Assumption 2 is preserved. In the second synthetic set ($\eta \in \{5, 10\}$), designed to violate Assumption 2, the mean minimum variance difference is -3.93—only slightly below zero and much closer to the baseline value (0.51) than to the perturbed setting (-39.42), indicating that Assumption 1 is largely preserved. Meanwhile, the mean maximum minimum gap increases to 4.42 (compared to 2.65 in the baseline), consistent with Assumption 2 being weakened, though the effect is considerably milder than in the semi-synthetic late-TTE subset.

In the semi-synthetic setting, Assumption 1 is violated in both subsets, $t_E < 816$ days and $t_E \geq 816$ days, since the variance of TTCs is lower than that of TTEs in both cases. However, Assumption 2 holds for $t_E < 816$, where the worst-case minimum TTC–TTE gap is 0 days, but is violated for $t_E \geq 816$, where this quantity increases to 93.97 days (**Table 9**).

Additional Results Table 10 shows the performance of CWITE on synthetic data where Assumption 1 is violated.

Table 10: Median MAE and 95% CIs for censored-like test individuals in the **synthetic-data setting with Assumption 1 violated** (25% cluster perturbation). Assumption 2 holds. Baselines do not use cluster assignments and are unaffected by the perturbation; full baseline results in **Table 1**.

Method	Censored-like Test Individuals
Standard Margin (best baseline)	7.01 [1.29, 12.93]
CWITE (Linear)	**3.64 [1.62, 7.43]**
CWITE (Exponential)	6.02 [1.82, 11.10]

C-Index Results We compute c-index values for censored-like and uncensored-like individuals by including all comparable pairs in which one member belongs to the subgroup (similar to censored or uncensored training data) and the other can be any individual in the test set. Tied predictions receive half credit. On synthetic data, c-index results closely mirror those based on MAE (**Table 11**). On real data, CWITE matches the best baseline on c-index for censored-like test individuals (0.76, 95% CI: [0.72, 0.80] vs. 0.76, [0.72, 0.79] for IPCW) while producing lower MAE, and performs comparably to baselines on uncensored-like test individuals (**Table 12**).

In the semi-synthetic setting, the pattern differs slightly. CWITE with exponential weighting performs best on censored-like test individuals with small true TTEs and on uncensored-like individuals with large true TTEs. In contrast, the Standard Margin method outperforms others on the remaining subgroups (**Table 16**). Thus, when optimizing for c-index in populations where Assumption 1 or Assumption 2 is violated, Standard Margin or CWITE with exponential weighting is recommended, depending on the characteristics of the subgroup of interest.

Varying λ results CWITE introduces two key parameters beyond those used in standard survival analysis methods: the weighting function $g(\cdot)$ and the scaling coefficient λ. For each censored data point, its loss is scaled by the ratio $g(t_{\text{obs}})/g(t_{\max})$, where $t_{\max}$ is the maximum observed censoring time within the cluster. The function $g(\cdot)$ thus acts as a *soft filter*, modulating the relative contribution of each censored point based on its proximity to the cluster's maximum. An exponential form of g assigns greater weight to censored points near $t_{\max}$, while a linear form yields a more uniform weighting. Separately, the coefficient λ is applied to all censored losses and controls their *overall contribution* to the objective function.

In our main experiments, we focus primarily on varying the weighting function $g(\cdot)$. Here, we also investigate the effect of the weighting coefficient λ. We find that λ has the greatest impact when CWITE is used with the exponential form of g, as it amplifies the effect of the exponential weights. In contrast, when using a linear g, model performance is largely insensitive to the choice of λ, suggesting that the linear weights exert a more uniform influence regardless of magnitude (**Table 17**). Thus, g should be chosen based on cluster reliability, while λ should be chosen based on the plausibility of close-to-event censoring and whether early-TTE accuracy is the main prediction priority.

Real-Data Stratification Diagnostics The logistic regression model used to estimate the probability of being uncensored in the real-data setting had moderate discrimination (AUROC: 0.632, 95% CI: [0.625, 0.639]) and imperfect calibration (Brier score: 0.248, 95% CI: [0.246, 0.250]). Because the model is used only to stratify test individuals by similarity to uncensored training support, ranking is the key requirement. This ranking was adequate for subgroup construction: observed uncensoring rates increased monotonically across predicted-propensity bins, from 17.3% to 67.0%.

Clustering Results Although CWITE is compatible with any clustering algorithm, we implement it with k-means, which can be unstable in high-dimensional settings. On the semi-synthetic dataset, where clusters are designed to be informative, k-means produced clusters that best satisfied Assumption 1, outperforming alternatives such as k-means with

Table 11: Median c-index and 95% CI for test individuals in the **synthetic-data setting** that are most similar to censored and uncensored training data.

Method	Censored-like Test Individuals	Uncensored-like Test Individuals
IPCW	0.52 [0.42, 0.76]	0.95 [0.54, 0.99]
Standard	0.53 [0.45, 0.78]	0.95 [0.50, 0.99]
Standard Margin	0.64 [0.44, 0.93]	0.95 [0.50, 0.99]
CWITE	**0.92 [0.67, 0.98]**	0.95 [0.64, 0.99]

Table 12: Median c-index and 95% CI for test individuals in the **real-data setting** that are most similar to censored and uncensored training data after feature selection. Values are computed with ties receiving half credit and with each subgroup ranked relative to the full test set.

Model	Censored-like Test Individuals	Uncensored-like Test Individuals
IPCW	**0.76** [0.72, 0.79]	**0.54** [**0.50, 0.58**]
Standard	0.61 [0.56, 0.65]	0.53 [0.49, 0.57]
Standard Margin	0.70 [0.68, 0.72]	0.52 [0.48, 0.56]
CWITE	**0.76** [**0.72, 0.80**]	0.53 [0.49, 0.57]

principal component analysis, Gaussian Mixture Models, and spectral clustering (see Appendix Table 13). Randomly perturbing synthetic cluster assignments increased CWITE's censored-like MAE from 1.89 to 3.64 days (Tables 1 and 10), showing that cluster quality matters when the underlying clusters are informative. In contrast, after feature selection in the real-data setting, CWITE was stable to clustering choices. Varying the number of k-means clusters from 2 to 40 changed censored-like MAE by at most 0.20 days. Relative to selected k-means, alternative clustering methods changed censored-like MAE only modestly: k-means with PCA increased MAE by 0.11 days, GMM increased MAE by 0.03 days, and spectral clustering increased MAE by 0.08 days (**Tables 14 and 15**). Replacing k-means with random clustering increased censored-like MAE by only 0.11 days, and using one global no-cluster group increased censored-like MAE by 0.12 days. These real-data sensitivity analyses suggest that, in this high-dimensional setting, CWITE's performance does not depend strongly on the specific learned clustering procedure.

Table 13: Mean difference on the entire semi-synthetic dataset between within-cluster TTC variance and TTE variance across clustering methods. Less negative values indicate better satisfaction of Assumption 1, with k-means achieving the smallest difference.

Dataset	k-means	k-means + PCA	GMM	Spectral Clustering
SUPPORT	-6.000	-41.292	-33.961	-44.996

Table 14: Clustering sensitivity in the **real-data setting**: median MAE with 95% CIs after feature selection.

Model	Censored-like Test Individuals	Uncensored-like Test Individuals
CWITE k-means selected	**6.50** [**5.55, 7.40**]	6.50 [5.54, 7.61]
CWITE random clusters	6.61 [5.68, 7.54]	6.52 [5.50, 7.71]
CWITE global no clusters	6.62 [5.68, 7.51]	6.41 [5.39, 7.58]
CWITE k-means rerun	**6.50** [**5.55, 7.40**]	6.50 [5.54, 7.61]
CWITE k-means + PCA	6.61 [5.67, 7.52]	**6.38** [**5.43, 7.48**]
CWITE GMM	6.53 [5.61, 7.44]	6.39 [5.43, 7.49]
CWITE spectral	6.58 [5.66, 7.51]	6.44 [5.46, 7.54]

Table 15: Clustering sensitivity in the **real-data setting**: median c-index with 95% CIs after feature selection.

Model	Censored-like Test Individuals	Uncensored-like Test Individuals
CWITE k-means selected	0.76 [0.72, **0.80**]	0.53 [0.49, 0.57]
CWITE random clusters	**0.77** [**0.73**, 0.80]	0.52 [0.47, 0.56]
CWITE global no clusters	**0.77** [**0.73**, 0.80]	0.53 [0.49, 0.57]
CWITE k-means rerun	0.76 [0.72, **0.80**]	0.53 [0.49, 0.57]
CWITE k-means + PCA	**0.77** [**0.73**, 0.80]	0.53 [0.50, 0.58]
CWITE GMM	**0.77** [**0.73**, 0.80]	**0.54** [**0.50, 0.58**]
CWITE spectral	**0.77** [**0.73**, 0.80]	0.53 [0.49, 0.57]

Table 16: Median c-index with 95% CIs for test individuals in the **semi-synthetic data setting**. Assumption 1 is always violated. Results are segmented by whether Assumption 2 holds (true TTE <816 days) or not (true TTE ≥816 days).

True TTE <816 days (Assumption 2 holds)		
Method	Censored-like Test Individuals	Uncensored-like
IPCW	0.611 [0.560, 0.649]	0.654 [0.613, 0.687]
Standard	0.562 [0.514, 0.607]	0.645 [0.624, 0.663]
Standard Margin	0.679 [0.545, 0.974]	**0.686 [0.522, 0.978]**
CWITE (Exponential)	**0.711 [0.626, 0.780]**	0.663 [0.586, 0.724]
CWITE (Linear)	0.658 [0.630, 0.682]	0.681 [0.639, 0.715]
True TTE ≥816 days (Assumption 2 violated)		
Method	Censored-like Test Individuals	Uncensored-like
IPCW	0.644 [0.605, 0.670]	0.658 [0.611, 0.698]
Standard	0.606 [0.566, 0.654]	0.490 [0.442, 0.543]
Standard Margin	**0.688 [0.565, 0.987]**	0.701 [0.540, 0.987]
CWITE (Exponential)	0.675 [0.628, 0.717]	**0.711 [0.634, 0.764]**
CWITE (Linear)	0.678 [0.652, 0.700]	0.689 [0.649, 0.730]

Table 17: Median MAE (in days) with 95% CIs for test individuals in the **semi-synthetic data setting** using CWITE with linear or exponential weighting, with small and large overall weight λ.

True TTE <816 days (Assumption 2 holds)		
Method	Censored-like Test Individuals	Uncensored-like
CWITE - Exponential, Small λ	591.72 [420.28, 795.09]	581.16 [365.49, 884.92]
CWITE - Exponential, Large λ	667.63 [484.32, 875.48]	932.22 [591.61, 1112.45]
CWITE - Linear, Small λ	368.52 [338.49, 399.05]	232.73 [170.93, 318.22]
CWITE - Linear, Large λ	366.31 [338.85, 395.05]	231.41 [178.82, 323.45]
True TTE ≥816 days (Assumption 2 violated)		
Method	Censored-like Test Individuals	Uncensored-like
CWITE - Exponential, Small λ	366.77 [303.75, 509.25]	309.03 [254.19, 384.64]
CWITE - Exponential, Large λ	378.34 [308.10, 537.22]	309.03 [257.90, 403.26]
CWITE - Linear, Small λ	536.77 [495.44, 574.41]	559.52 [486.90, 636.62]
CWITE - Linear, Large λ	549.90 [512.34, 589.25]	577.31 [509.13, 651.44]

Table 18: Percentage of uncensored individuals in the real-data test set for whom each approach overpredicted TTE. Overprediction means that the predicted TTE was later than the true TTE, such that spontaneous labor occurred earlier than predicted. In a scheduling context, this could result in a planned induction being scheduled after spontaneous labor had already begun.

IPCW	Standard	Standard Margin	CWITE
49.4 [48.5, 50.1]	69.0 [68.2, 69.8]	54.4 [53.6, 55.2]	**45.3 [44.5, 46.1]**