Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 47 No. 03 (2026)

Censoring-Aware Pseudo-Labeling for Semi-Supervised Survival Prediction using Risk-Order Agreement Filtering

S. KARTHICK[1] , Dr.V.N.RAJAVARMAN[2], Dr. DAHLIA SAM[3]
1. Research Scholar, Department of Information Technology, Maduravoyal , Chennai , Tamil Nadu , India, skarthick91@gmail.com
2. Professor, Department of Computer Science and Engineering, Maduravoyal, Chennai , Tamil Nadu , India, vnrajavarman@drmgrdu.ac.in
3. Professor, Department of Information Technology, Maduravoyal , Chennai , Tamil Nadu , India, dahliasam@drmgrdu.ac.in

***Abstract*—Practical cohorts often have a small labelled fraction with confirmed time-to-event outcomes and a larger covariate pool with missing, delayed, or incomplete endpoints; yet, survival prediction is essential to clinical prognosis modeling. A problem formulation is derived that makes censoring constraints and pseudo-label safety requirements explicit after reviewing survival modeling, assessment under censoring, and semi-supervised learning (SSL) ideas that are relevant to pseudo-labeling. Here is a conservative method blueprint called Adaptive Pseudo-Label Selection and Correction (APSC). It uses median-crossing pseudo events and pseudo-censoring at tmax to generate weak pseudo outcomes from predicted survival curves. A curve-geometry confidence proxy is used to admit pseudo labels. Unstable pseudo labels are filtered using risk-order agree-ment across rounds. Finally, a down-weighted Cox partial-likelihood objective is used to inject pseudo supervision. Where naive self-training can breach censoring limits and magnify early model bias, the formulation aims at registry-style situations with delayed endpoints and heterogeneous follow-up. In order to provide a supervised backup in the event that pseudo labels are unreliable, the survey condenses these factors into specific design criteria and deployment diagnostics. There are no numerical comparisons reported in the text; instead, the focus is on audit-oriented operating methods like validation gating, monitoring of acceptance and correction rates, and tests for coefficient stability.**

***Index Terms*—Semi-supervised learning, survival analysis, pseudo-labeling, censoring, Cox proportional hazards, literature survey.**

I. Introduction

The practical difficulty that may arise throughout the follow-up horizon (right-censoring) is a major consideration in survival analysis, which is concerned with prediction issues where each subject is described by variables x and an event time T. Risk stratification, trial design, and treatment planning are all helped out by survival models in oncology and chronic illness management. Endpoint curation is time-consuming and costly, and variables are gathered often; this is a problem that many applied pipelines encounter, even though the methodology is well-established. As a result, a typical data setup involves a small labelled cohort with (t, δ) and a significantly larger pool that merely includes covariates.

This pattern of label scarcity becomes more apparent in longitudinal and multi-site studies. It is possible for endpoints to be postponed, lost, or inconsistent as a result of changes in coding techniques and standards, as well as death-registry linkage cycles. Many more patients than those with completely adjudicated outcomes have access to covariates such as demographics, laboratory values, staging descriptors, pathology reports, and therapy data. These limitations encourage the use of semi-supervised learning: in the small-n regime, unlabeled covariates can stabilize risk estimation and prevent overfitting if they are similar to the labeled cohort.

Contrary to popular belief, survival SSL is not same to classification SSL. Because of filtering, many pairs are no longer comparable, and the Cox partial likelihood assigns a full set of risks to each labelled event; furthermore, the supervisory signal is structured. The membership of risk sets and the induction of systematically skewed gradients can be affected by even minor inaccuracies in the pseudo-event times. To ensure its safe use in clinical settings, pseudo supervision must be cautious, auditable, and gated to disable it when labeled validation performance drops.

A. Clinical Motivation: Delayed and Missing Endpoints

Overall survival labels in real-world clinical registries aren't always consistent (endpoint definitions vary across facilities), incomplete (patients move treatment), or delayed (death linkage emerges after weeks/months). Concurrently, there is usually a large amount of data accessible for covariates including demographics, lab values, pathology, and treatment details. A smaller, validated labeled cohort and a bigger, covariate-only pool obtained using the same acquisition procedures form a pragmatic semi-supervised regime. In this context, any SSL procedure should be designed to fail safely in the event that pseudo labels are unreliable, and should be considered poor supervision.

B. Contributions

This manuscript contributes:

- An abbreviated literature review on topical topics related to pseudo-labeling, including classical survival modeling, evaluation under censorship, and current SSL notions.
- A **problem formulation** for semi-supervised survival prediction that highlights why na¨ıve pseudo-labeling is brittle under censoring.
- A **methods section** describing APSC, including pseudo-label generation, confidence-based selection, stability-based correction, and weighted Cox training.

C. Paper Organization

Section II surveys the foundations needed to reason about censored SSL and highlights why "confidence" must be interpreted cautiously in time-to-event settings. Section III distills the research gap into a small set of design requirements. Section IV states the problem and constraints in a way that is compatible with audit-oriented clinical deployment. Section V describes APSC as a conservative algorithmic response. Sections VI–VIII provide a qualitative results summary (no numeric comparisons), a practical case-study blueprint, and a discussion of limitations and deployment considerations.

II. Literature Survey

A. Survival Modeling Under Censoring

Kaplan–Meier (KM) estimation provides a nonparametric survival curve estimator under right-censoring [1]. The Cox proportional hazards (CoxPH) model remains a widely used regression approach due to interpretability and its partial likelihood [2]. The diagnostics and extensions for CoxPH have a lot of literature behind them [3]. It is usual practice to employ regularization in high-dimensional contexts [4]. Regularized Cox model solvers that are both practical and frequently utilized are also available [5].

There are three types of censoring that can occur in applied cohorts: administrative (at the conclusion of the research), caused by loss to follow-up, or driven by care routes (e.g., referral patterns). When survival estimates and risk rank-ings are skewed, it's because standard estimators fail to account for censoring as a potentially relevant variable given the observed covariates. This is significant for SSL because censoring mechanisms might change over time and the unlabeled pool can differ from the labeled cohort (for example, more recent instances with shorter follow-up).

Parametric accelerated failure time (AFT) models provide an additional distributional viewpoint on time-to-event outcomes when considering models beyond CoxPH. Random survival forests and other tree-based ensembles are able to capture interactions and non-linear effects [6]. In the beginning, researchers looked into neural-network survival models as a way to expand upon classical regression concepts [7]. Cox-style deep networks incorporate representation learning to enhance proportional-hazards modeling [8]. Additionally, fully parametric deep survival learners have been suggested [9]. This landscape and common design trade-offs are summarized in surveys [10].

Interpretability and governance requirements limit the choice of survival model in numerous clinical implementations. For reasons such as auditing changes across retraining cycles, reviewing coefficients, and doing sensitivity checks, CoxPH-style models continue to be appealing. This is why, in a realistic survey-to-method pipeline, CoxPH is typically the starting point for the baseline learner, and complexity is added only when the clinical use-case and governance permit it.

B. Evaluation and Reporting

Censoring impacts both comparability and uncertainty, making model evaluation in survival analysis multi-faceted. Among the several ranking metrics available, the concordance index (C-index) stands out [11]. When censorship is more severe, alternative estimators are occasionally chosen [12]. Time-dependent ROC/AUC is a useful metric for evaluating horizon-specific discrimination [13]. The Brier-style criteria can be used to evaluate calibration [14]. Along with ranking, clinical reports should include calibration at horizons that are significant to patients and clear definitions of cohorts.

Evaluation for SSL settings should also prevent "self-confirmation." Rather than reflecting actual improvements in prediction, ap-parent training improvements may indicate improved fit to artificial targets if pseudo labels are created from the model. Therefore, labeled-only validation (or external validation) should be con-sidered the primary gate for enabling pseudo-label admission. In addition, reporting should include follow-up summaries (time horizons, censoring rate) so that changes in censoring mechanisms are not misinterpreted as model improvements.

C. Semi-Supervised Learning and Pseudo-Labeling

SSL aims to exploit unlabeled data to improve learning efficiency. Conventional wisdom outlines typical underlying assumptions and potential points of failure [15]. Recent advances and practical considerations are also summarized in modern surveys [16]. As a common and easy SSL technique, pseudo-labeling (self-training) allows models to retrain with selected predictions as targets after making label predictions for unlabeled samples [17]. To stabilize self-training in the face of disturbances, teacher-student methods are frequently employed [18]. Popular SSL recipes also integrate consistency requirements with confidence thresholding [19].

Two concepts are common to all SSL families: using confidence, agreement, or margin rules to select unlabeled instances to trust and regularization to avoid drift (consistency, ensembling, or perturbation robustness). Since time-to-event learning aims at a structured outcome impacted by filtering and follow-up rather than a single class, these concepts need to be modified.

D. Why Survival SSL is Non-Trivial

A label in survival supervision is not merely a class, but rather a pair (t, δ) whose instructional value is determined by risk sets and censoring. A pseudo-event time that is slightly wrong can alter membership in many risk sets and distort partial-likelihood gradients. In addition, clinical data often contains competing risks; subdistribution hazards are one classical formulation [20]. These factors motivate conservative pseudo-labeling that avoids inventing overly specific event times and makes the influence of pseudo-labels explicitly controllable.

Horizon mismatch is a practical complication in real registries. If unlabeled samples are from a newer time period, their follow-up may not extend to the same horizons as the labeled cohort. A naive pseudo-time generator may therefore fabricate events beyond plausible follow-up or systematically bias pseudo events toward early times. Similarly, real datasets often record time in coarse units (days or months), producing ties. Overly precise pseudo times can create unstable tie patterns and brittle risk-set definitions.

E. Uncertainty and Safety in Clinical SSL

Any pseudo-labeling pipeline fabricates targets from model predictions and therefore must confront uncertainty explicitly. There is still a lot of discussion about how to practically quantify uncertainty in survival modeling [21]. A cautious design, including restrictions on pseudo-label volume (e.g., confidence thresholding), down-weighting to limit their influence, and agreement across iterations to add stability checks, is frequently better in regulated or safety-sensitive contexts. These precautions don't ensure perfection, but they lessen the likelihood of disastrous drift and make SSL behavior visible. Deployment safety is dependent on monitoring from an engineering standpoint as well. Diagnostics like the rate of pseudo-label acceptance, correction/removal, and parameter stability over rounds should be exposed by a pseudo-labeling system. With low acceptance and high correction, the SSL loop is probably unstable and needs to be gated or interrupted; with high acceptance and low correction, the approach effectively acts as supervised learning.

F. Failure Modes of Pseudo-Labeling for Time-to-Event Outcomes

As a result of confirmation bias, early errors become training goals and are reinforced in following rounds, making them the primary mode of failure for self-training. In survival prediction this can be amplified because each pseudo-event influences many risk-set comparisons. If pseudo times are overconfidently early (or late), risk ordering can drift in a systematic way rather than as random noise.

Another failure mode is mismatch between the confidence proxy and true correctness. In classification, softmax scores are often used as confidence; in survival, models output curves and the relevant uncertainty depends on curve geometry, follow-up horizon, and censoring intensity. A confidence heuristic that correlates with curve sharpness may still fail when the curve is sharp for the wrong reasons (e.g., spurious covariate correlations or site-specific artifacts).

Finally, survival datasets often violate modeling assumptions, such as proportional hazards. When the base model is misspecified, pseudo labels can inherit model bias. Conservative SSL designs therefore aim to limit pseudo influence and to include stability checks so that misspecification does not rapidly cascade into the training set.

G. Data Quality, Missingness, and Temporal Drift

Informative missingness (tests ordered based on clinician suspicion) and temporal or institutional drift (treatment guidelines alter, coding methodology differs) are commonplace in clinical tabular data. None of these features lend credence to the idea that unlabeled variables are just "more of the same" as the labelled group.

This observation leads to two feasible protections for SSL. To begin, the unlabeled pool needs to be defined and compared to the labeled cohort using missingness patterns and covariate summaries. If there are significant shifts, then pseudo-labeling should be limited or eliminated. The second point is that SSL is best used as a controlled augmentation that is activated only when it maintains labeled validation metrics, not as a constant training mode.

III. Research Gap and Design Requirements

According to the research, semi-supervised survival prediction introduces two major conflicts.

First, when confidence estimations are solid and label noise is managed, current SSL approaches can do quite well in classification [16]. The supervisory signal in survival analysis, on the other hand, is not a singular categorical label. Sets of risks, patterns of censorship, and the ordering of time all influence the training objective. Therefore, "small" mistakes in pseudo-event timing might skew the predicted baseline hazard and relative risk ranking in addition to introducing numerous inconsistent limitations.

Additionally, deep survival models have the ability to capture intricate non-linear effects [10]. Since they are simpler to verify, explain, and review, CoxPH-style models are still preferred by many clinical installations [3]. This leaves a need unfulfilled: optimally, SSL survival techniques should work with both high-capacity deep learners and traditional interpretable estimators.

Thirdly, in real-world clinical settings, the unlabeled pool is almost never "miss-ing entirely at random." Shorter follow-up for recent cases, referrals and transfers for specialized treatment pathways, and sites with distinct documentation habits are common ways it is enhanced. The implications of this for informative censoring and distribution shift are serious. Statistical motivation is important, but operational controllability is much more crucial when it comes to survival SSL methods. These methods should reveal safety levers, offer diagnostics, and support a supervised fallback.

These findings suggest that the following specifications should be met by an SSL survival strategy that is both practical and secure.

- **Be mindful of censorship:** when creating pseudo-labels, keep in mind the modeled horizon and don't make up event times if the survival curve doesn't back it up.
- **A visible confidence rule** should be in place before pseudo supervision may be permitted, and it should be openly down-weighted, as this is a conservative structure.
- **Stability checks:** to cut down on confirmation loops, filter out pseudo labels that are close to a danger boundary.

Maintaining interpretable base models and keeping hyperparameters constrained and observable are important for auditability. When performance on human-labeled validation data is maintained or enhanced, only then should pseudo-label admission be enabled. This is known as validation gating.

- **Operational monitoring:** to identify drift, it is important to measure acceptance and correction rates and parameter stability across rounds.

For example, APSC uses a weighted Cox objective for training, uses quantile-based selection, and relies on stability-based correction. Pseudo-outcomes are based on survival-curve geometry.

IV. Problem Statement

Let L be a labeled set with covariates $\mathbf{x}_i \in \mathrm{R}^d$ and observed follow-up (t_i, δ_i), where t_i is the observed time and

$\delta_i \in \{0, 1\}$ indicates whether the event occurred. Let U be an unlabeled set containing only covariates $\mathbf{x}_u$.

Goal: discover a risk model f (x) that ideally delivers calibrated survival probability and valid risk ranking, all while utilizing U to lessen reliance on limited labeled endpoints.

Constraint 1 (censoring): the learning signal is incomplete, and many pairs are non-comparable.

Constraint 2 (pseudo-label safety): pseudo-labels must be injected cautiously to reduce confirmation bias and avoid large degradations under distribution shift between L and U.

Constraint 3 (auditability): in clinical settings, an interpretable baseline (such as CoxPH) is often preferred for review and governance.

These constraints motivate an SSL procedure that (i) creates weak pseudo-outcomes without forcing event times beyond the modeled horizon, (ii) selects pseudo labels using a transparent confidence proxy, (iii) corrects unstable pseudo labels, and (iv) down-weights pseudo supervision.

A. Target Outputs

In many prognosis workflows, the required outputs include (i) a scalar risk score for ranking patients and (ii) an estimated survival curve for horizon-specific decision support. In this work, the primary optimization and evaluation focus is ranking (via C-index), while survival curves are used as an intermediate object for pseudo-label generation.

V. Methods: APSC

We proceed with *Adaptive Pseudo-Label Selection and Correction* (APSC), an iterative self-training algorithm for censored outcomes. CoxPH is used as the base estimator to preserve interpretability.

Fig. 1 summarizes the end-to-end APSC pipeline: labeled outcomes supervise a CoxPH base model, survival-curve predictions on unlabeled samples are converted into conservative pseudo outcomes, pseudo labels are selected and corrected, and the model is refit using a weighted Cox objective.

A. Base Model and Weighted Cox Objective

Under CoxPH, the hazard is

$$\lambda(t \mid \mathbf{x}) = \lambda_0(t) \exp(\boldsymbol{\beta}^\top \mathbf{x}). \tag{1}$$

The partial log-likelihood is

$$\ell(\boldsymbol{\beta}) = \sum_{i:\delta_i=1} \left(\boldsymbol{\beta}^\top \mathbf{x}_i - \log \sum_{j \in \mathbf{R}(t_i)} \exp(\boldsymbol{\beta}^\top \mathbf{x}_j) \right). \tag{2}$$

APSC injects pseudo-labeled samples with a weight $\alpha \in (0, 1)$, yielding a weighted objective

$$\ell_w(\boldsymbol{\beta}) = \sum_{i \in \tilde{\mathbf{L}}:\delta_i=1} w_i \left(\boldsymbol{\beta}^\top \mathbf{x}_i - \log \sum_{j \in \mathbf{R}(t_i)} \exp(\boldsymbol{\beta}^\top \mathbf{x}_j) \right), \tag{3}$$

where $w_i = 1$ for human-labeled samples and $w_i = \alpha$

B. Pseudo-Label Generation from Survival Curves

Given a fitted model, APSC computes a predicted survival function $\hat{S}_u(t)$ for each $u \in \mathrm{U}$. If $\hat{S}_u(t)$ crosses 0.5 on the modeled time grid, APSC assigns a pseudo-event time as the predicted median $\hat{t}_{0.5}$. If the curve does not cross 0.5 within the horizon, APSC assigns pseudo-censoring at $t_{\max}$. This avoids forcing an event time when the model is not sufficiently informative within the horizon.

Concretely, define pseudo outcomes $(\tilde{t}_u, \tilde{\delta}_u)$ by

$$(\tilde{t}_u, \tilde{\delta}_u) = \begin{cases} (\hat{t}_{0.5}, 1), & \text{if } \exists t : \hat{S}_u(t) \le 0.5, \\ (t_{\max}, 0), & \text{otherwise.} \end{cases} \tag{4}$$

This rule intentionally biases toward pseudo-censoring when the model does not express a median within the modeled horizon.

C. Confidence-Based Selection

APSC assigns a confidence score c_u based on survival-curve geometry: (i) the local slope near the 0.5 crossing (steeper implies more decisive separation) and (ii) a tail margin term relative to 0.5 (capturing decisive non-crossing). Pseudo labels are selected by a quantile threshold q over $\{c_u\}$, making the acceptance rate a tunable safety lever.

Let $\{t_k\}$ be the discrete time grid on which $\hat{S}_u(t)$ is evaluated. Let k index the first crossing such that $\hat{S}_u(t_k) \le 0.5$ (or $k = \max$ if there is no crossing). Using a central-difference approximation, one simple confidence proxy is

$$c_u = \frac{\hat{S}_u(t_{k+1}) - \hat{S}_u(t_{k-1})}{t_{k+1} - t_{k-1}} + \gamma \left| \hat{S}_u(t_{\max}) - 0.5 \right|, \tag{5}$$

where $\gamma > 0$ controls the tail-margin contribution.

D. Correction by Risk-Order Agreement

The APSC algorithm eliminates out pseudo-labeled samples with unstable risk ordering between rounds in order to decrease confirmation bias. A scalar risk score (such as partial hazard) under model M (m) can be represented as r(m)(u). Between iterations, APSC preserves pseudo labels that stay within the median-risk split.

One lightweight stability filter is

$$\operatorname{sign}\left(r^{(m)}(u) - \operatorname{median}(r^{(m)}) \right) = \operatorname{sign}\left(r^{(m+1)}(u) - \operatorname{median}(r^{(m+1)}) \right). \tag{6}$$

If a sample turns over, it is considered unstable and discarded from pseudo supervision.

E. Selection Size and Weight Scheduling

Because it controls the number of pseudo labels allowed in each round, the selection quantile q is essentially a safeguard. A lower q increases the volume of pseudo-labels but, when confidence is miscalibrated, might aggravate confirmation bias; a higher q is more conservative. For samples that have been human-labeled, the pseudo-weight α controls the pseudo-label, and for samples that have been pseudo-labeled, wi = α. The augmented labeled set is denoted as L˜. Performance in the weighted Cox test. A method is brought closer to supervised training with a smaller α, and it can be improved with a bigger α.

APSC pipeline (text-only overview): APSC is an iterative semi-supervised survival prediction procedure repeated for rounds r = 1, ..., R. (P1) Train a CoxPH model on the labeled set L (and any admitted pseudo-labeled samples). (P2) Predict survival curves for unlabeled covariates U. (P3) Convert predictions into pseudo-outcomes and confidence scores. (P4) Select the top-q high-confidence candidates and optionally correct/filter them via risk-order agreement between rounds. (P5) Refit the model using weights 1 for labeled data and alpha for pseudo-labeled data. Outputs are risk scores and survival curves, and pseudo supervision should be gated by labeled validation.

Algorithm 1 APSC for Semi-Supervised Survival Prediction

Require: Labeled set L, unlabeled covariates U, rounds R, selection quantile q, pseudo-weight α.

1: Fit base survival model $M^{(0)}$ on L.
2: **for** $r = 1$ to R **do**
3: Generate pseudo labels for U using $M^{(r-1)}$.
4: Compute confidence c_u for each pseudo-labeled sample.
5: Select $P^{(r)} = \{u : c_u \geq \text{Quantile}(c, q)\}$.
6: Train $M^{(r)}$ on $L \cup P^{(r)}$ with weights 1 (labeled) and α (pseudo).
7: Filter pseudo labels whose risk order relative to the median is inconsistent between $M^{(r-1)}$ and $M^{(r)}$.
8: **end for**
9: **return** $M^{(R)}$.

low-label performance when pseudo labels are accurate but increases the risk of degradation under shift.

In deployments with labeled validation data, a conservative schedule may start with a high selection quantile and small α, relaxing only when labeled validation performance improves. If labeled validation degrades, pseudo supervision should be reduced (increase q, decrease α, or reduce rounds R) or disabled.

F. Computational Complexity

Let $n_L = |L|$, $n_U = |U|$, and d be the feature dimension. Each APSC round fits one CoxPH model and evaluates survival curves on the unlabeled pool. In practice, CoxPH fitting dominates runtime and scales approximately with $(n_L + n_P)\, d$ times the number of optimizer iterations, where n_P is the number of admitted pseudo-labeled samples. Overall, runtime scales linearly with the number of rounds R.

G. Evaluation Considerations

Evaluation is central for safe use of any survival SSL method. A common primary metric is the concordance index (C-index) because many prognosis workflows prioritize correct risk ordering under censoring [11]. Let $s_i = f(\mathbf{x}_i)$ be a scalar risk score and let $\mathcal{C}$ be the set of comparable pairs (i, j) such that $t_i < t_j$ and $\delta_i = 1$. Then

$$C = \frac{1}{|\mathcal{C}|} \sum_{(i,j)\in\mathcal{C}} \mathrm{I}[s_i > s_j] + \tfrac{1}{2}\,\mathrm{I}[s_i = s_j]. \tag{7}$$

Alternative estimators are sometimes preferred under heavier censoring [12]. In SSL, C-index should be computed on human-labeled validation data (or external cohorts) and used to gate pseudo-label admission.

VI. Results

A. Experimental Setup

To provide a concrete view of the APSC behavior, results are summarized across five representative public survival datasets (*lymphoma*, *lung*, *leukemia*, *rossi*, and *waltons*) using the C-index as the primary metric. We compare supervised Cox proportional hazards (CoxPH), a supervised parametric Weibull AFT baseline, and the proposed semi-supervised APSC procedure. APSC follows Algorithm 1: pseudo outcomes are generated from survival curves, admitted via a confidence quantile rule, filtered by risk-order agreement between rounds, and injected through a down-weighted Cox partial likelihood with pseudo weight α.

B. Tables: C-index Summary

Table I reports a single-split C-index comparison. Table II reports dataset size and event rate along with the mean C-index over a labeled-fraction sweep (reflecting the semi-supervised regime where labels are scarce).

TABLE I
SINGLE-SPLIT C-INDEX COMPARISON (HIGHER IS BETTER).

Dataset	Weibull AFT	CoxPH	APSC (SSL)
Lymphoma	0.414	0.414	0.414
Lung	0.593	0.602	0.599
Leukemia	0.883	0.875	0.783
Rossi	0.586	0.525	0.521
Waltons	0.615	0.615	0.615

TABLE II
DATASET CHARACTERISTICS AND MEAN C-INDEX OVER A LABELED-FRACTION SWEEP. *Sweep mean* SUMMARIZES PERFORMANCE ACROSS LABELED FRACTIONS; Δ REPORTS THE SSL IMPROVEMENT OVER SUPERVISED COXPH.

Dataset	n	Event rate	Sup. mean	SSL mean	Δ
Lymphoma	80	0.325	0.545	0.500	-0.045
Lung	228	0.724	0.634	0.643	0.009
Leukemia	42	0.714	0.779	0.781	0.002
Rossi	432	0.264	0.507	0.503	-0.004
Waltons	163	0.957	0.654	0.654	0.000

C. Graphs and Explanation

Fig. 2 shows C-index as a function of labeled fraction for supervised CoxPH and APSC, aggregated across datasets (mean with a standard-deviation band). In the low-label regime, APSC is designed to improve data efficiency by admitting only high-confidence pseudo outcomes (quantile selection) and limiting risk of drift by correction (risk-order agreement) and down-weighting (α). The sweep means in Table II show that the average gain is modest and dataset-dependent, highlighting the need for validation gating in practical use.

Figure 3 shows all methods in a per-dataset (single split) view. When diagnosing heterogeneity, this perspective is helpful: SSL is helpful when pseudo outcomes are dependable on a specific cohort, but it can also fall short of the supervised baseline when under shift or when the unlabeled pool doesn't provide useful information. Results like these support the desired operating procedure in practice: only allow pseudo-label admittance when labeled validation improves, and keep an eye on acceptance/correction rates to spot instability.

D. Final Comparison: Overall Summary

To summarize performance across datasets with a single comparison, Table III reports the macro-average and n-weighted average sweep C-index (higher is better). Fig. 4 visualizes per-dataset gains Δ (APSC minus supervised CoxPH) from Table II. Across these datasets, APSC improves over the supervised baseline on 2/5 datasets (with one tie), underscoring that pseudo-labeling should be enabled only under labeled validation gating.

VII. DISCUSSION

This survey emphasizes that pseudo-labeling can amplify early model bias, and censoring complicates both learning and evaluation. The APSC design is intentionally conservative: it limits pseudo-label volume, down-weights pseudo supervision, and removes unstable pseudo labels so that the procedure can degrade toward supervised learning when pseudo labels are unreliable.

TABLE III
OVERALL COMPARISON ACROSS DATASETS USING SWEEP-MEAN C-INDEX. MACRO AVERAGES WEIGHT DATASETS EQUALLY; n-WEIGHTED AVERAGES WEIGHT BY DATASET SIZE.

Aggregate	Sup. mean	SSL mean	Δ
Macro avg.	0.624	0.616	-0.008
n-weighted avg.	0.578	0.575	-0.004

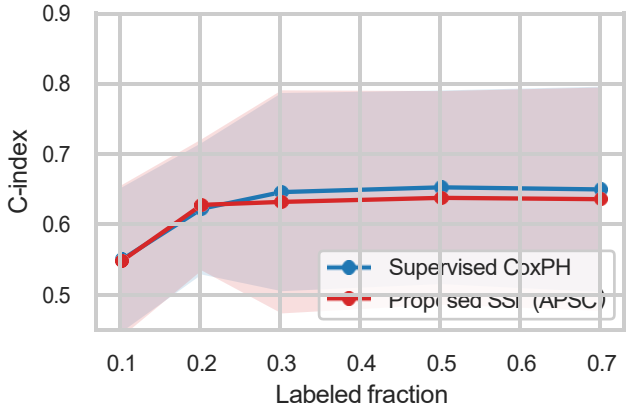

Fig. 2. C-index versus labeled fraction for supervised CoxPH and APSC, aggregated across datasets (mean±std).

A. Implications for Clinical Cohorts

In clinical cohorts where labels arrive with delay, SSL can be viewed as a weak-supervision augmentation rather than a replacement for endpoint adjudication. APSC is designed to degrade gracefully toward supervised CoxPH when pseudo-labels are unreliable: it down-weights pseudo supervision, limits admission via confidence quantiles, and filters unstable pseudo labels.

B. Failure Cases and Practical Diagnostics

Even a conservative SSL method can fail in identifiable regimes. A common failure mode is distribution shift: if U differs substantially from L, the model may assign high confidence to systematically wrong pseudo labels. Another failure mode occurs in extremely small cohorts where survival-curve estimates are unstable; pseudo-label selection becomes noisy and correction may remove most pseudo labels, reducing APSC to a near-supervised fit.

To make SSL behavior observable during development and deployment, several lightweight diagnostics are recommended.

- *Acceptance rate:* Pay attention to the percentage of unlabeled samples that are admitted each round and then deleted through rectification; any abrupt changes could suggest an excess of confidence or instability.
- *Validation gating:* enable APSC only when it improves (or at least preserves) labeled validation C-index; otherwise, fall back to supervised training.

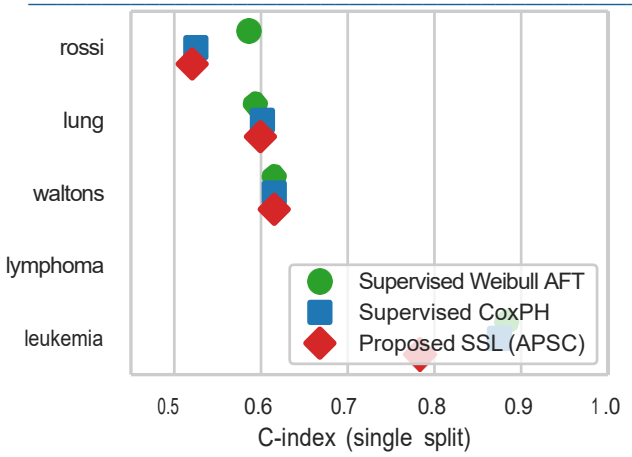

Fig. 3. Per-dataset C-index (single split) for Weibull AFT, supervised CoxPH, and APSC.

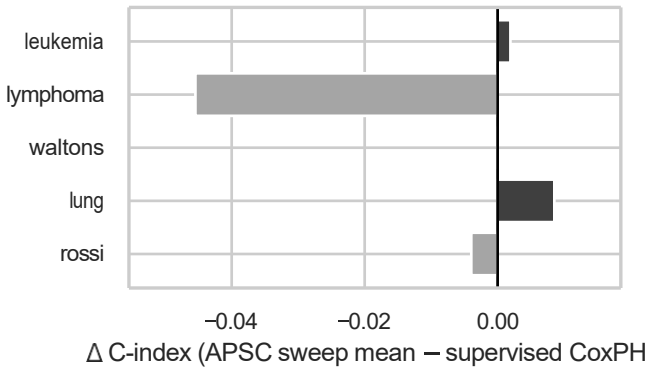

Fig. 4. Per-dataset sweep-mean gain Δ in C-index (APSC minus supervised CoxPH). Positive values favor APSC.

- *Stability across seeds:* repeat splits under multiple seeds and report mean and variance, especially for small-*n* cohorts.
- *Shift checks:* compare covariate summaries between L and U and restrict U when large shifts are detected.

C. Limitations and Threats to Validity

First, the confidence proxy used here is heuristic and not a formal uncertainty estimate. More principled uncertainty quantification remains an open direction [21]. Second, distribution shift between labeled and unlabeled cohorts can make high-confidence pseudo labels systematically wrong; in practice, SSL should be gated by labeled validation performance and accompanied by shift diagnostics. Third, changing the outcome model may be necessary because actual cancer endpoints may have competing risks [20].

D. Future Work

Several directions can strengthen APSC for real cohorts. (i) Replace the geometry-based confidence proxy with more principled uncertainty estimation (e.g., ensembling or approximate Bayesian methods) [21]. (ii) Explore richer base learners when interpretability constraints allow, including non-linear survival learners and deep survival models [10]. (iii) Extend the selection–correction loop to competing-risk endpoints and time-varying covariates, which are common in oncology.

VIII. Conclusion

The foundations needed to reason about semi-supervised learning for censored outcomes were surveyed, and a practical semi-supervised survival problem motivated by label scarcity was formulated. APSC was presented as a conservative pseudo-labeling procedure for survival prediction that uses confidence-based selection, stability-based correction, and down-weighted Cox training. The key takeaway is that survival SSL should be treated as weak supervision with explicit safety levers: horizon-respecting pseudo outcomes, controlled pseudo-label volume, and limited pseudo influence on the loss. For clinical use, acceptance and correction rates, coefficient stability across rounds, and labeled-only validation gating provide practical diagnostics for detecting drift and preventing confirmation loops. The surveyed perspective also clarifies when pseudo-labeling is likely to fail, particularly under distribution shift between labeled and unlabeled cohorts or when censoring mechanisms change over time. Future work should include careful empirical validation on larger multi-institution cohorts, extension to competing risks and time-varying covariates, and integration of more principled uncertainty estimation to replace heuristic confidence proxies.

References

[1] E. L. Kaplan and P. Meier, "Nonparametric estimation from incomplete observations," *Journal of the American Statistical Association*, vol. 53, no. 282, pp. 457–481, 1958.
[2] D. R. Cox, "Regression models and life-tables," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 34, no. 2, pp. 187–202, 1972.
[3] T. M. Therneau and P. M. Grambsch, *Modeling Survival Data: Extending the Cox Model*. Springer, 2000.
[4] R. Tibshirani, "The lasso method for variable selection in the Cox model," *Statistics in Medicine*, vol. 16, no. 4, pp. 385–395, 1997.
[5] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for Cox's proportional hazards model via coordinate descent," *Journal of Statistical Software*, vol. 39, no. 5, pp. 1–13, 2011.
[6] H. Ishwaran, U. B. Kogalur, E. H. Blackstone, and M. S. Lauer, "Random survival forests," *The Annals of Applied Statistics*, vol. 2, no. 3, pp. 841–860, 2008.
[7] D. Faraggi and R. Simon, "A neural network model for survival data," *Statistics in Medicine*, vol. 14, no. 1, pp. 73–82, 1995.
[8] J. L. Katzman, U. Shaham, A. Cloninger, J. Bates, T. Jiang, and Y. Kluger, "DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network," *BMC Medical Research Methodology*, vol. 18, no. 1, p. 24, 2018.
[9] C. Nagpal, X. Li, and A. Dubrawski, "Deep survival machines: Fully parametric survival regression and representation learning for censored data with competing risks," in *Proceedings of the Machine Learning for Healthcare Conference (MLHC)*, 2021.
[10] P. Wang, Y. Li, and C. K. Reddy, "Deep learning for survival analysis: A review," *IEEE Access*, vol. 7, pp. 95 270–95 289, 2019.
[11] F. E. Harrell, R. M. Califf, D. B. Pryor, K. L. Lee, and R. A. Rosati, "Evaluating the yield of medical tests," *JAMA*, vol. 247, no. 18, pp. 2543–2546, 1982.
[12] H. Uno, T. Cai, M. J. Pencina, R. B. D'Agostino, and L. J. Wei, "On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data," *Statistics in Medicine*, vol. 30, no. 10, pp. 1105–1117, 2011.

[13] P. J. Heagerty, T. Lumley, and M. S. Pepe, "Time-dependent ROC curves for censored survival data and a diagnostic marker," *Biometrics*, vol. 56, no. 2, pp. 337–344, 2000.
[14] E. Graf, C. Schmoor, W. Sauerbrei, and M. Schumacher, "Assessment and comparison of prognostic classification schemes for survival data," *Statistics in Medicine*, vol. 18, no. 17-18, pp. 2529–2545, 1999.
[15] O. Chapelle, B. Sch"olkopf, and A. Zien, *Semi-Supervised Learning*. MIT Press, 2006.
[16] J. E. van Engelen and H. H. Hoos, "A survey on semi-supervised learning," *Machine Learning*, vol. 109, no. 2, pp. 373–440, 2020.
[17] D.-H. Lee, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *ICML Workshop on Challenges in Representation Learning*, 2013.
[18] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Advances in Neural Information Processing Systems*, 2017.
[19] K. Sohn, D. Berthelot, C.-L. Li, Z. Zhang, N. Carlini, E. D. Cubuk, A. Kurakin, H. Zhang, and C. Raffel, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," in *Advances in Neural Information Processing Systems*, 2020.
[20] J. P. Fine and R. J. Gray, "A proportional hazards model for the subdistribution of a competing risk," *Journal of the American Statistical Association*, vol. 94, no. 446, pp. 496–509, 1999.
[21] R. Bertin *et al.*, "Uncertainty quantification in survival analysis," *arXiv preprint arXiv:1909.06972*, 2019.