Final Report: Reproducing Auton-Survival for Survival Analysis with Censored Data

(Nagpal, Potosnak, and Dubrawski, 2022)

Ningyuan Xie

Master of Computer Science
Siebel School of Computing and Data Science
University of Illinois at Urbana-Champaign
nxie3@illinois.edu

CS598 Deep Learning for Healthcare

Outline

1. General Problem: Survival Analysis with Censored Data

Survival Analysis in Healthcare:

- **Challenge:** Predict time-to-event outcomes (e.g., mortality) when some patients have incomplete follow-up
- **Censoring:** Patients may leave study before event occurs $\rightarrow$ we only know they survived until censoring time

Traditional Methods Fall Short:

- Cox Proportional Hazards assumes proportional hazards (too restrictive)
- Cannot model heterogeneous patient populations
- Limited counterfactual estimation and no unified framework

Gap: Existing tools (scikit-survival, lifelines) lack integrated deep learning, phenotyping, and counterfactual methods

Outline

2. Paper's Approach: Auton-Survival Framework

Three Integrated Components:

1. **Survival Regression Models**
   - **Deep Cox PH:** Neural network extends Cox model
   - **Deep Survival Machines (DSM):** Mixture of parametric distributions

$$S(t|X) = \sum_{k=1}^{K} \pi_k(X) S_k(t|\mu_k(X), \sigma_k(X))$$

   - **Deep Cox Mixtures (DCM):** Combines Cox with mixture components
2. **Phenotyping Methods:** Intersectional, unsupervised, supervised (DCM), counterfactual (CMHE)
3. **Censoring-Adjusted Evaluation:** IPCW for Brier Score, AUC, Concordance

Key Innovation: Mixture components relax proportional hazards assumption, enabling flexible modeling of heterogeneous populations and supervised phenotyping

Outline

3. Reproduction Results: Scope & Performance

Successfully Reproduced: 15/15 Core Functionalities

Survival Regression (6):

- Deep Cox PH
- SurvivalModel Wrapper
- SurvivalRegressionCV
- Importance Weighting
- Counterfactual Regression
- Time-Varying Regression

Phenotyping (6):

- Intersectional
- Unsupervised Clustering
- Supervised (DCM)
- Phenotype Purity
- Virtual Twins
- CMHE Counterfactual

Evaluation (3):

- Censoring-Adjusted Metrics
- RMST Treatment Effect
- Propensity-Adjusted

Metric	30d	90d	180d	IBS
Brier Score	0.2013	0.2282	0.2429	0.2272
TD-AUC	0.5781	0.6068	0.6263	–
C-Index	0.5561	0.5703	0.5808	–

Results Comparison:

- Paper reports pipeline only (no numerical baselines)
- My reproduction matched the pipeline: same data, models, APIs
- Differences were expected due to stochastic sampling and updated environment
- Qualitative behavior aligned with authors' descriptions

Outline

4. Extension Studies: Mixture Component Ablation

Question: Do mixture components affect and improve performance?

Model	k	Concordance	AUC	Time (s)
DSM	1	0.395	0.652	0.6
	2	0.378	0.671	0.7
	3	0.365	0.678	0.7
	5	0.352	0.683	0.8
DCM	1	0.384	0.641	0.6
	2	**0.487**	**0.698**	**0.7**
	3	0.465	0.691	0.7
	5	0.451	0.685	0.8

Key Findings:

- **DSM:** Showed diminishing returns, potential overfitting with more components
- **DCM:** Showed significant improvement with k=2 (+26.77% concordance vs k=1)
- **Conclusion:** Validated authors' hypothesis—mixtures capture heterogeneity

4. Extension Studies: Architecture Depth Ablation

Question: Do deeper architectures affect and improve performance?

Model	Architecture	Concordance	Parameters
DeepCoxPH	[100]	0.351	1.0x
	[100, 100]	0.375 (+6.88%)	3.59x
DSM	[100]	0.378	1.0x
	[100, 100]	0.383 (+1.36%)	3.35x
DCM	[100]	**0.487**	1.0x
	[100, 100]	**0.398 (-18.29%)**	3.54x

Key Findings:

- **DeepCoxPH:** Showed modest improvement (+6.88%) with 3.59x parameters
- **DSM:** Showed marginal improvement (+1.36%)
- **DCM:** Degradation (-18.29%) suggested overfitting
- **Conclusion:** Shallow architectures sufficed; mixture components already captured complexity

4. Extension Studies: Cross-Dataset Validation

Question: Can models trained on one dataset generalize to another?

Experimental Design:

- Train on SUPPORT (500 samples, 38 features) — critically ill patients
- Test on PBC (312 samples, 25 features) — liver disease patients
- Extreme domain shift: different features, populations, outcomes

Cross-Dataset Results

Model	Train	Test	Result
DeepCoxPH	SUPPORT	PBC	No overlap
DSM	SUPPORT	PBC	No overlap
DCM	SUPPORT	PBC	No overlap

Key Findings:

- **All models generated** "no valid horizons" status
- SUPPORT and PBC have **zero overlapping features**
- Revealed fundamental limitation: models are **feature-dependent**
- Clinical deployment requires **domain-specific training**