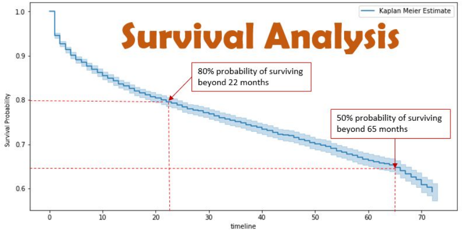
Revisit to Deep Cox Mixtures for Survival Regression

CS598 - Deep Learning for Healthcare

Chris Yu hmyu2@illinois.edu
Jimmy Lee jl279@illinois.edu

Introduction


Survival Analysis

80% probability of surviving beyond 22 months

50% probability of surviving beyond 65 months

What is Survival Analysis?

- Statistical methods for analyzing the expected duration of time until one or more events happen.
- Commonly used in healthcare, engineering, economics, etc.
- Deals with "censored data" (where the event of interest hasn't occurred by the end of the study).

Problems Statements

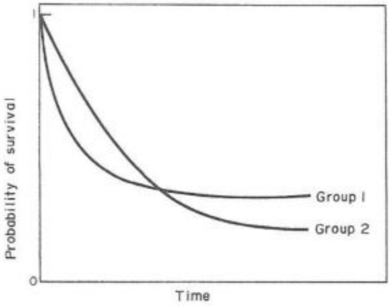


Figure 2: Survival curves that don't satisfy the proportional hazards assumption.

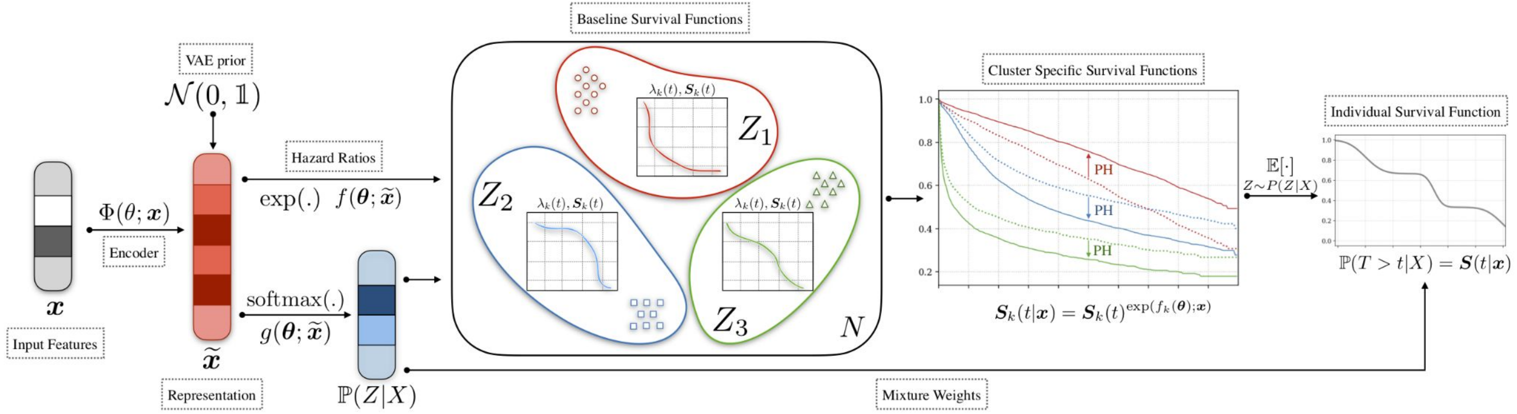
Widely-Used Models: Cox Proportional Hazards
Major Problem: Poor Calibration

Deep Learning Models: DeepSurv, DeepHit,Deep Survival Machines (DSM).
Problem: Emphasis on *ranking performance* rather than the *absolute score values*.

Real World Consequences overestimating risk for millions, impacting healthcare decisions.

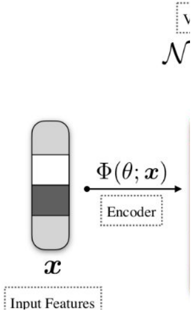
Deep Cox Mixtures (DCM): Concept

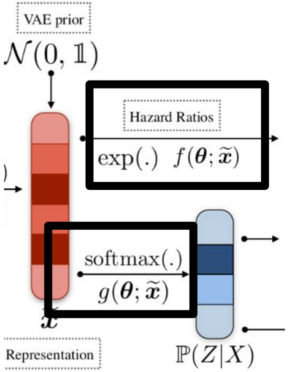
- **Model Idea:** Extends the standard Cox model to handle diverse patient risk profiles by modeling survival as a finite **mixture of K distinct Cox models**.
- Individuals are probabilistically assigned to **latent subgroups** $Z \in \{1, ..., K\}$.
- Each subgroup has its own baseline survival curve $S_k(t)$, estimated **non-parametrically**.
- The final predicted survival $S(t|x)$ for an individual is a weighted average of the cluster-specific survival functions, based on their subgroup probabilities $P(Z|X = x)$.



VAE prior
$\mathcal{N}(0, \mathbb{1})$
$\boldsymbol{x}$
Input Features
$\Phi(\theta; \boldsymbol{x})$
Encoder
$\widetilde{\boldsymbol{x}}$
Representation
Hazard Ratios
$\exp(.)$ $f(\boldsymbol{\theta}; \widetilde{\boldsymbol{x}})$
softmax(.)
$g(\boldsymbol{\theta}; \widetilde{\boldsymbol{x}})$
$\mathbb{P}(Z|X)$
Baseline Survival Functions
$\lambda_k(t), \boldsymbol{S}_k(t)$
Z_1
Z_2
Z_3
N
Cluster Specific Survival Functions
1.0
0.8
0.6
0.4
0.2
PH
PH
PH
$\boldsymbol{S}_k(t|\boldsymbol{x}) = \boldsymbol{S}_k(t)^{\exp(f_k(\boldsymbol{\theta});\boldsymbol{x})}$
$\mathbb{E}[\cdot]$
$Z \sim P(Z|X)$
Individual Survival Function
1.0
0.8
0.6
0.4
0.2
0.0
$\mathbb{P}(T > t|X) = \boldsymbol{S}(t|\boldsymbol{x})$
Mixture Weights

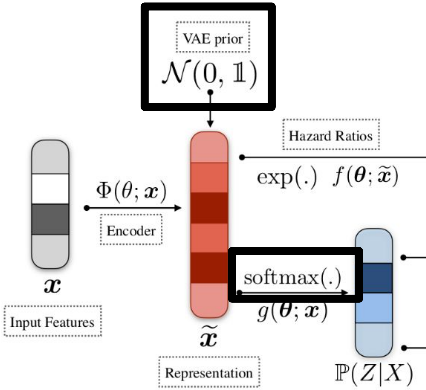
Deep Cox Mixtures (DCM): Concept




VAE prior
$\mathcal{N}(0, \mathbb{1})$
Hazard Ratios
$\exp(.)$ $f(\boldsymbol{\theta}; \widetilde{\boldsymbol{x}})$
softmax(.)
$g(\boldsymbol{\theta}; \widetilde{\boldsymbol{x}})$
$\widetilde{\boldsymbol{x}}$
Representation
$\mathbb{P}(Z|X)$

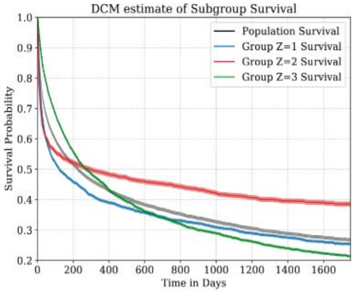
- **Input Data:** Uses standard **right-censored survival data**: $D = \{(x_i, \delta_i, u_i)\}_{i=1}^{N}$, where x_i are covariates, δ_i is the event indicator (1=event, 0=censored), and u_i is the observed time.
- **Learning Method:** Model parameters are optimized using **Maximum Likelihood Estimation (MLE)**.

$$\mathcal{L}(\theta, \Lambda_k) = \prod_{i=1}^{|\mathcal{D}|} \int_Z (\lambda_k(u_i|x_i))^{\delta_i} S_k(u_i|x_i) \mathbb{P}(Z = k|x_i)$$

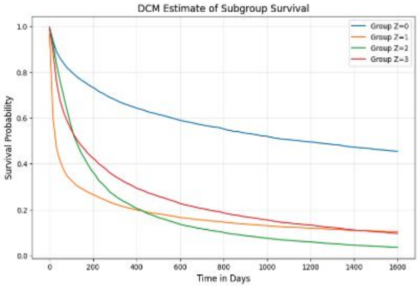

VAE prior
$\mathcal{N}(0, \mathbb{1})$
Hazard Ratios
$\exp(.)$ $f(\boldsymbol{\theta}; \widetilde{\boldsymbol{x}})$
$\Phi(\theta; \boldsymbol{x})$
Encoder
softmax(.)
$g(\boldsymbol{\theta}; \boldsymbol{x})$
$\boldsymbol{x}$
Input Features
$\widetilde{\boldsymbol{x}}$
Representation
$\mathbb{P}(Z|X)$

Reproduction: Kaplan-Meier Curves

By simulating survival curves for the SUPPORT dataset based on our DCM implementation, we observed curves with a **similar shape** to those presented in the original paper

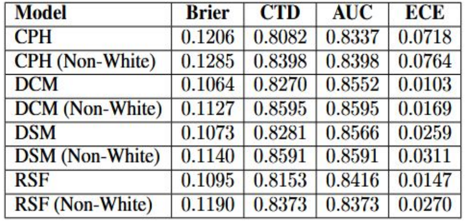


Original paper



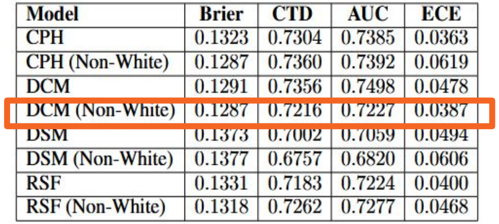
Our Reproduction

Reproduction: Metrics Comparison SEER



Model	Brier	CTD	AUC	ECE
CPH	0.1206	0.8082	0.8337	0.0718
CPH (Non-White)	0.1285	0.8398	0.8398	0.0764
DCM	0.1064	0.8270	0.8552	0.0103
DCM (Non-White)	0.1127	0.8595	0.8595	0.0169
DSM	0.1073	0.8281	0.8566	0.0259
DSM (Non-White)	0.1140	0.8591	0.8591	0.0311
RSF	0.1095	0.8153	0.8416	0.0147
RSF (Non-White)	0.1190	0.8373	0.8373	0.0270

Original paper

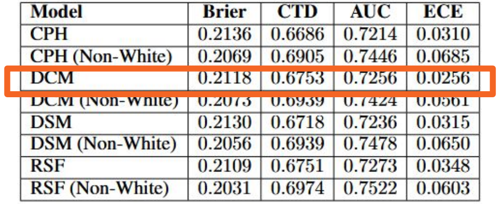


Model	Brier	CTD	AUC	ECE
CPH	0.1323	0.7304	0.7385	0.0363
CPH (Non-White)	0.1287	0.7360	0.7392	0.0619
DCM	0.1291	0.7356	0.7498	0.0478
DCM (Non-White)	0.1287	0.7216	0.7227	0.0387
DSM	0.1373	0.7002	0.7059	0.0494
DSM (Non-White)	0.1377	0.6757	0.6820	0.0606
RSF	0.1331	0.7183	0.7224	0.0400
RSF (Non-White)	0.1318	0.7262	0.7277	0.0468

Our Reproduction

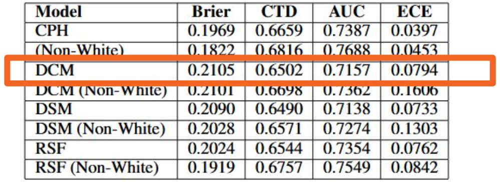
- Lower ECE in the minority group indicates better calibration, which is a **positive sign** for DCM's calibration on this dataset compared to the original paper's claims.
- Our **CTD (0.7356)** was close to the original DCM indicating similar ranking ability.

Reproduction: Metrics Comparison SUPPORT



Model	Brier	CTD	AUC	ECE
CPH	0.2136	0.6686	0.7214	0.0310
CPH (Non-White)	0.2069	0.6905	0.7446	0.0685
DCM	0.2118	0.6753	0.7256	0.0256
DCM (Non-White)	0.2073	0.6939	0.7424	0.0561
DSM	0.2130	0.6718	0.7236	0.0315
DSM (Non-White)	0.2056	0.6939	0.7478	0.0650
RSF	0.2109	0.6751	0.7273	0.0348
RSF (Non-White)	0.2031	0.6974	0.7522	0.0603

Original paper



Model	Brier	CTD	AUC	ECE
CPH	0.1969	0.6659	0.7387	0.0397
(Non-White)	0.1822	0.6816	0.7688	0.0453
DCM	0.2105	0.6502	0.7157	0.0794
DCM (Non-White)	0.2101	0.6698	0.7362	0.1606
DSM	0.2090	0.6490	0.7138	0.0733
DSM (Non-White)	0.2028	0.6571	0.7274	0.1303
RSF	0.2024	0.6544	0.7354	0.0762
RSF (Non-White)	0.1919	0.6757	0.7549	0.0842

Our Reproduction

- Our DCM reproduction showed **weaker calibration (significantly higher ECE: 0.0794)** and **lower ranking (CTD: 0.6502)** compared to the original paper's reported DCM metrics (ECE: 0.0256, CTD: 0.6753).

Reasons for Observed Differences

Overall Reproducibility: The original paper was found to be **partially reproducible**.

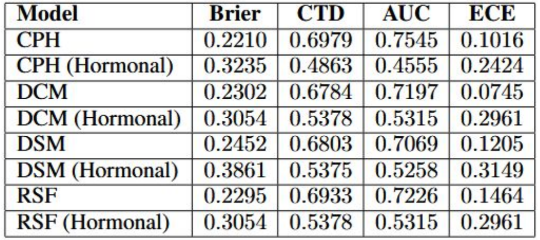
Factors Limiting Exact Reproduction:

- Unavailable **original training code**.
- **Unclear step of training and evaluation.**

Evaluating DCM on GBSG Dataset

- The German Breast Cancer Study Group (GBSG)
- **Dataset Origin:** a late 1980s clinical trial for node-positive breast cancer cases.
- **Data:** Contains complete prognostic profiles for 686 patients, focusing on disease and treatment responses.

GBSG Dataset : Performances



Model	Brier	CTD	AUC	ECE
CPH	0.2210	0.6979	0.7545	0.1016
CPH (Hormonal)	0.3235	0.4863	0.4555	0.2424
DCM	0.2302	0.6784	0.7197	0.0745
DCM (Hormonal)	0.3054	0.5378	0.5315	0.2961
DSM	0.2452	0.6803	0.7069	0.1205
DSM (Hormonal)	0.3861	0.5375	0.5258	0.3149
RSF	0.2295	0.6933	0.7226	0.1464
RSF (Hormonal)	0.3054	0.5378	0.5315	0.2961

- Our DCM model achieved **competitive performance** compared to other deep learning approaches (DSM/RSF).
- DCM showed **great performance in ranking accuracy (CTD 0.6784)** compared to the traditional CPH model (CTD 0.6979).
- DCM maintained **reasonable calibration (ECE 0.0745)** on this dataset.

Conclusion

Survival models struggle with **calibration** and **non-proportional hazards**.

DCM is promising: uses **deep learning** and **latent subgroups** for survival modeling.

Reproduction study results were **mixed**:

- Successfully reproduced **Kaplan-Meier curve shapes**.
- Metrics varied; **calibration (ECE) challenging** on SUPPORT.
- Strong performance on **SEER** (better than baselines).
- Competitive performance on **GBSG** (good ranking, reasonable calibration).

Reproducibility was **partial**: limited by missing code, unclear steps, preprocessing details.