Let the Experts Speak: Improving Survival Prediction & Calibration via Mixture-of-Experts Heads

Todd Morrill TODD@CS.COLUMBIA.EDU
Department of Computer Science, Columbia University, USA
Aahlad Puli AAHLAD@NYU.EDU
Department of Computer Science, New York University, USA
Murad Megjhani MM5025@CUMC.COLUMBIA.EDU
Department of Neurology, Columbia University Medical Center, USA
Department of Computer Science, Barnard College, USA
Soojin Park SP3291@CUMC.COLUMBIA.EDU
Department of Neurology, Columbia University Medical Center, USA
Department of Biomedical Informatics, Columbia University Medical Center, USA
NewYork-Presbyterian Hospital at Columbia University Medical Center, USA
Richard Zemel ZEMEL@CS.COLUMBIA.EDU
Department of Computer Science, Columbia University, USA

Abstract

Deep mixture-of-experts models have attracted a lot of attention for survival analysis problems, particularly for their ability to cluster similar patients together. In practice, grouping often comes at the expense of key metrics such calibration error and predictive accuracy. This is due to the restrictive inductive bias that mixture-of-experts imposes, that predictions for individual patients must look like predictions for the group they're assigned to. Might we be able to discover patient group structure, where it exists, while *improving* calibration and predictive accuracy? In this work, we introduce several discrete-time deep mixture-of-experts (MoE) based architectures for survival analysis problems, one of which achieves all desiderata: clustering, calibration, and predictive accuracy. We show that a key differentiator between this array of MoEs is how expressive their experts are. We find that more expressive experts that tailor predictions per patient outperform experts that rely on fixed group prototypes.

Keywords: survival analysis, mixture-of-experts, clustering, calibration, accuracy

Data and Code Availability In this work, we use three publicly available datasets. SUPPORT2 is a freely downloadable survival analysis dataset (Connors et al., 1995).[1] The Sepsis dataset (Reyna et al., 2020) is also freely available after a training course.[2] MNIST is a well-known research dataset that can be downloaded from a variety of sources.[3] Our Github code repository is available at https://github.com/ToddMorrill/survival-moe.

Institutional Review Board (IRB) Our research does not require IRB approval.

1. Introduction

AI has the potential to have a profound impact on clinical decision support systems (CDSS). AI systems are being developed for medical imaging (Erickson et al., 2017), disease diagnosis (Ahsan et al., 2022), and many other clinical support settings. However, AI for CDSS faces barriers to adoption due to clinician mistrust of model predictions (Eltawil et al., 2023). In this work, we focus on what clinicians care about most: highly accurate models, where probabilities have intuitive meaning (i.e., calibration), and interpretability of the model's decision-making process–in this case, the ability to reason by analogy to similar patients. Our work addresses survival analysis, where the task is to predict when clinical events will occur (i.e., time-to-event regression) while con-

1. https://archive.ics.uci.edu/dataset/880/support2

2. https://physionet.org/content/challenge-2019/1.0.0/

3. https://docs.pytorch.org/vision/main/generated/torchvision.datasets.MNIST.html

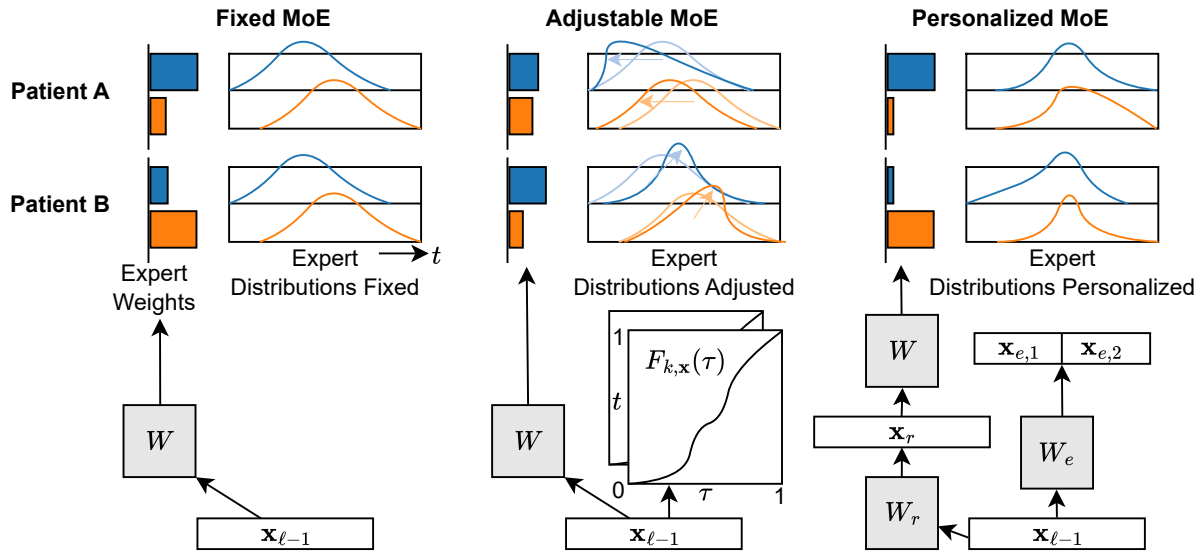


Figure 1: Illustration of our three proposed mixture-of-experts (MoE) architectures for survival analysis. Left - **Fixed MoE** showing the same expert distributions for patients A and B with differing expert weights. Middle - **Adjustable MoE** where fixed expert distributions (faint color) are adjusted per patient (dark color). Right - **Personalized MoE** where all experts produce a custom event distribution for all patients.

tending with right-censoring–not observing the event time for a subset of the patients.

Mixture-of-experts (MoE) models for medical survival analysis are particularly appealing for the above desiderata due to their ability to discover latent groups of patients. MoEs are defined by two key components: (i) a router that assigns patients to groups and (ii) a set of experts that produce event distributions for each group (Nagpal et al., 2021b; Hou et al., 2023; Buginga and e Silva, 2024). Our goal is to investigate how expert expressivity in deep, discrete-time MoE heads affects calibration and predictive accuracy under matched capacity, which has not been carefully studied in the medical survival modeling setting. In our study, experts range from fixed prototypes to per-patient parameterizations. We therefore compare three MoE heads that differ only in expert expressivity to isolate its impact and include standard non-MoE baselines for context.

The first architecture, **Fixed MoE**, uses several experts where each expert learns an associated event distribution, which is then fixed across all patients. The second architecture, **Adjustable MoE**, again learns a prototypical event distribution per expert, but can be adjusted to the individual patient. The third architecture, **Personalized MoE**, uses experts that each form a custom event distribution for individual patients. Only the third architecture is able to cluster patients *and* outperform strong baseline models on calibration, concordance index, and time-dependent Brier score. The contributions of our work are as follows: (i) we introduce three discrete-time deep MoE based survival architectures, one of which achieves excellent clustering, calibration error, concordance index, and time-dependent Brier score and (ii) we report that the expressiveness of the experts is a key differentiator between discrete-time deep MoE-based survival analysis models, supported by a targeted set of experiments on these three architectures.

2. Related Work

Building survival models with flexible model classes, such as Gaussian processes (Alaa and van der Schaar, 2017) or neural networks (Ranganath et al., 2016), and rich distribution families (Miscouridou et al., 2018), yields better predictions of risk than standard linear Cox or accelerated failure-time (AFT) mod-

els. One can retain the structure of AFT models but improve prediction by using deep representations of the covariates (Zhong et al., 2021; Norman et al., 2024). Further generalizations that retain structure were introduced as conditional transformation models (Hothorn et al., 2018; Campanella et al., 2022) These choices improve expressivity to allow for better approximations of complicated survival distributions but do not reveal any structure in the data.

To better understand structure in the data, prior works (Manduchi et al., 2022; Nagpal et al., 2021b,a) cluster data, with each cluster assigned an "expert" distribution that is structured as Weibull-like or Cox-Proportional-Hazards-like curves. Buginga and e Silva (2024) instead work with unconditional Kaplan-Meier curves for each cluster assignment. However, clustering while restricting the expert distribution structure or using input-independent (unconditional) experts can hurt metrics like concordance (Hou et al., 2023; Buginga and e Silva, 2024). While Manduchi et al. (2022); Nagpal et al. (2021b,a) do report improved metrics and clustering, fitting their models involves running variational inference techniques which require more work than is standard for using default supervised learning pipelines. We work with discrete-time categorical distributions to remove limitations from Cox-like structures and study varying degrees of conditioning of experts to build calibrated and accurate survival models, without introducing additional complications to standard training. While prior work has addressed discrete-time survival modeling (Yu et al., 2011; Fotso, 2018; Lee et al., 2018; Kvamme and Borgan, 2021), we are aware of no work addressing discrete-time deep MoE models for survival analysis.

There are similarities between MoE methods and ensemble-based approaches, particularly random survival forests (Ishwaran et al., 2008). Both frameworks rely on aggregating/ensembling predictions from multiple specialized components (experts or trees). Recalibration is another notable line of work that aims to improve calibration in survival models via post-hoc methods (e.g., conformal prediction) (Qi et al., 2024).

3. Methods

We now describe our three proposed deep mixture-of-experts (MoE) based survival architectures. All models are feedforward deep learning models that use information from a patient's record (e.g., demographic data, physiological data, etc.) to forecast when they are likely to have a clinical event. Raw patient records $\mathbf{x}_0^i \in \mathbb{R}^a$ for patients $i = 1 \dots N$ and a the number of raw features are composed of categorical indicators (e.g., gender, etc.), and standardized continuous features (e.g., heart rate), meaning that for continuous features we subtract the mean and divide by the standard deviation. We learn embeddings for categorical indicators. After embedding lookups are concatenated with continuous features, this results in a d-dimensional feature vector that is fed to the model. Let $\mathbf{x} = \mathbf{x}_{\ell-1}^i \in \mathbb{R}^h$ be shorthand for the h-dimensional penultimate hidden state representation of our feedforward model with ℓ layers for the i^{th} patient. We will now describe the ℓ^{th} layer (i.e., the final layer), which is the MoE head in all architectures. All methods are trained using a discrete-time Multitask Logistic Regression (MTLR) style loss function (Yu et al., 2011; Fotso, 2018) that predicts a monotone label sequence ("once event occurs, it stays on"). This loss function has been shown to be well-calibrated (Haider et al., 2020). The loss functions are described in Appendix Section B.

3.1. Fixed Mixture-of-Experts

The Fixed MoE architecture is representative of a class of prior works that use learned, but fixed per-patient, event distributions (Hou et al., 2023). The Fixed MoE is composed of a learnable router $W \in \mathbb{R}^{n \times h}$ where n is the number of experts and a collection of learnable experts $M \in \mathbb{R}^{n \times m}$, where each row can be mapped to a distribution over the m possible discrete event times. We produce a probability mass function (PMF) $\mathbf{p}$ over discrete event times as follows

$$\boldsymbol{\alpha}(\mathbf{x}) = \text{softmax}(\mathbf{x}W^\intercal/\kappa) \tag{1}$$

$$\mathbf{p} = \boldsymbol{\alpha}M' = \sum_{j=1}^{n} \alpha_j M'_j \tag{2}$$

where α_j is the weight on expert j, M'_j is the j^{th} row of the parameter matrix M after it has been normalized to form a discrete event distribution, and κ is a learnable temperature parameter that modulates the sharpness of expert selection. To our knowledge, prior discrete-time neural survival models (e.g., DeepHit (Lee et al., 2018)) do not use an explicit MoE head over a categorical time grid, and MoE-style survival models have largely been continuous-time parametric or nonparametric (Hou et al., 2023; Nagpal et al., 2021b).

3.2. Adjustable Mixture-of-Experts

The adjustable MoE can be seen as an exemplar from a class of models that transform a prototypical event distribution per expert to form a custom event distribution per patient (Nagpal et al., 2021b; Campanella et al., 2022; Manduchi et al., 2022). The adjustable MoE learns an event distribution per expert and then *warps* it per patient. The following equations describe in detail how we warp prototypical event distributions but the important point is that with relatively few additional parameters per expert, we can flexibly adjust the event distributions per patient.

Let $\mathbf{t} \in [0,1]^m$ denote the canonical grid over the m discrete time bins, with $t_j = j/(m-1)$ for $j = 0, \ldots, m-1$. Each expert k maintains a prototype vector $M_k \in \mathbb{R}^m$ of unnormalized scores over event times. To tailor expert k to patient i, we define a strictly monotone bijection between the expert's internal time $\boldsymbol{\tau} \in [0,1]^m$ and the canonical grid $\mathbf{t}$:

$$\underbrace{\phi_{k,\mathbf{x}} : \boldsymbol{\tau} \to \mathbf{t}}_{\text{forward}}, \qquad \underbrace{\psi_{k,\mathbf{x}} : \mathbf{t} \to \boldsymbol{\tau}}_{\text{inverse}} = \phi_{k,\mathbf{x}}^{-1}.$$

Concretely, we take the forward map to be a normalized mixture of $r = 2$ logistic cumulative density functions (CDF),

$$F_{k,\mathbf{x}}(u) = \sum_{r=1}^{2} w_{k,r}(\mathbf{x})\, \sigma(a_{k,r}(\mathbf{x})\,[\,u - c_{k,r}(\mathbf{x})\,]), \quad (3)$$

$$\sigma(z) = \frac{1}{1+e^{-z}}, \quad (4)$$

Weights $w_{k,r}(\mathbf{x}) > 0$ with $\sum_r w_{k,r}(\mathbf{x}) = 1$ enforce a linear combination of the two logistic CDFs that results in a function $F_{k,\mathbf{x}}(u)$ with a maximum value of 1. Slopes $a_{k,r}(\mathbf{x}) > 0$ control the steepness of each logistic component. Ordered centers $0 < c_{k,1}(\mathbf{x}) < c_{k,2}(\mathbf{x}) < 1$ control the location of each logistic component and we enforce an ordering on the centers so that the second logistic CDF is always to the right of the first. Importantly, $w_{k,r}(\mathbf{x})$, $a_{k,r}(\mathbf{x})$, and $c_{k,r}(\mathbf{x})$ are all learned linear functions of the final hidden state representation $\mathbf{x}$, allowing each patient to have a custom warping function per expert. We chose $r = 2$ to allow for a more flexible warping of the event-time distribution compared to a single logistic CDF–see Figure 1 for an example of the function's shape with $r = 2$. Note that the left endpoint $F_{k,\mathbf{x}}(0)$ and right endpoint $F_{k,\mathbf{x}}(1)$ are not guaranteed to be 0 and 1, respectively, based on the parameterization in Equation 3. We therefore enforce a mapping from $[0,1] \to [0,1]$ via endpoint normalization,

$$\tilde{F}_{k,\mathbf{x}}(u) \;=\; \frac{F_{k,\mathbf{x}}(u) - F_{k,\mathbf{x}}(0)}{F_{k,\mathbf{x}}(1) - F_{k,\mathbf{x}}(0)} \in [0,1], \quad (5)$$

and define $\phi_{k,\mathbf{x}}(u) = \tilde{F}_{k,\mathbf{x}}(u)$ to be the forward map and $\psi_{k,\mathbf{x}} = \phi_{k,\mathbf{x}}^{-1}$ to be the inverse map. Given a canonical gridpoint $t_j \in [0,1]$, we compute $u_{k,j} = (m-1)\,\psi_{k,\mathbf{x}}(t_j)$, which is an approximate index into the m scores stored in M_k on the model's internal time grid. We linearly interpolate the scores with

$$i_0 = \lfloor u_{k,j} \rfloor, \quad (6)$$

$$i_1 = \min(i_0 + 1, m-1), \quad (7)$$

$$w_{k,j} = u_{k,j} - i_0, \quad (8)$$

$$\tilde{M}_{k,j} = (1 - w_{k,j})\, M_{k,i_0} + w_{k,j}\, M_{k,i_1}. \quad (9)$$

In practice, we evaluate $\psi_{k,\mathbf{x}}(t_j)$ via a bisection solver (see Appendix Section H for more details).

Let $\tilde{M}'_k$ denote the row-wise normalization of $\tilde{M}_k$ into an event-time PMF. The final per-patient PMF is then

$$\boldsymbol{\alpha}(\mathbf{x}) = \text{softmax}(\mathbf{x}W^\intercal/\kappa), \quad (10)$$

$$\mathbf{p} = \boldsymbol{\alpha}\tilde{M}' = \sum_{k=1}^{n} \alpha_k(\mathbf{x})\, \tilde{M}'_k. \quad (11)$$

This transformation family generalizes simple shift/scale warps and can emulate proportional-hazards style tilts while allowing richer early/late adjustments with minimal additional parameters per expert (Zhong et al., 2021; Nagpal et al., 2021b; Campanella et al., 2022; Manduchi et al., 2022).

3.3. Personalized Mixture-of-Experts

The Personalized MoE architecture belongs to a collection of models that provide the most flexibility to the experts to form custom event distributions per patient (e.g., Deep Survival Machines (Nagpal et al., 2021a)). Since this architecture generates new expert distributions for each patient, we first project the final hidden state representaion $\mathbf{x}$ to a router representation with $\mathbf{x}_r = \mathbf{x}W_r^\intercal$ for $W_r \in \mathbb{R}^{h \times h}$ and an expert representation with $\mathbf{x}_e = \mathbf{x}W_e^\intercal$ for $W_e \in \mathbb{R}^{h \times h}$. We then divide the expert representation into n evenly-sized chunks $\mathbf{x}_{e,k}$ for $k = 1, \ldots, n$, which are each fed to a linear layer denoted $L_k \in \mathbb{R}^{m \times (h/n)}$ to form an event-distribution to obtain $M_k(\mathbf{x}_{e,k}) = \mathbf{x}_{e,k}L_k^\intercal$, which collectively form the dynamic matrix of unnormalized densities over event times $M(\mathbf{x}_e) \in \mathbb{R}^{n \times m}$.

The final PMF is then

$$\boldsymbol{\alpha}(\mathbf{x}_r) = \text{softmax}(\mathbf{x}_r W^\intercal / \kappa) \tag{12}$$

$$\mathbf{p} = \boldsymbol{\alpha} M(\mathbf{x}_e)' = \sum_{j=1}^{n} \alpha_j M(\mathbf{x}_e)'_j \tag{13}$$

where $M(\mathbf{x}_e)'_j$ is the j^{th} row of the parameter matrix $M(\mathbf{x}_e)$ after it has been normalized to form a discrete event distribution.

Our method is notable for its parameter efficiency due to chunking the expert representation and may force the model to use independent information for each expert, which may be beneficial for clustering and predictive accuracy.

4. Experiments

We experiment with 3 datasets to probe the properties and capabilities of our proposed methods. Survival MNIST is a synthetic dataset (see Figure 2), which allows us to probe the models' abilities to predict event times and recover latent groups. We censor 15% of examples. SUPPORT2 is a survival analysis dataset with 9,105 examples and ∼32% censoring (Connors et al., 1995). Sepsis is a larger dataset with 40,336 patient records and only 2,932 positive instances of sepsis, making this a very challenging anomaly detection survival analysis task (Reyna et al., 2020). For the Sepsis dataset, the first 100 hours of each patient's ICU stay was summarized to perform a retrospective prediction, helpful when using partially labeled historical datasets to accelerate human labeling (Singer et al., 2016; Henry et al., 2019). Patients in the Sepsis dataset were administratively censored after 100 hours.

Performance Assessment As a comprehensive set of metrics covering performance with respect to accuracy and uncertainty calibration, we measure Harrell's concordance index (Harrell et al., 1996), the time-dependent Brier score (standard inverse probability of censoring weighting (IPCW) (Gerds and Schumacher, 2006)) at 3 key points in time: the 25th, 50th, and 75th percentile of time bins, and equal mass expected calibration error (ECE) (Roelofs et al., 2022) adjusted with IPCW and averaged over all time bins. To control for parameter count as a potential source of performance differences, we ensure that all neural models have the same number of parameters (see Appendix Table C). All measurements are averaged over 5 random seeds. In order to obtain a ranking of the models by performance, *for each random seed* we take the difference between the MoE model's metric and MTLR model's metric, and then average that over the 5 runs to obtain an average performance gap, which is reported in parentheses in Table 1. Network architectures and hyperparameters are described in Appendix Section C. This experiment is important for two reasons: (i) it establishes calibration and predictive accuracy metrics for common non-clustering baseline models such as Cox proportional hazards (CoxPH) (Cox, 1972), random survival forests (RSF) (Ishwaran et al., 2008), and MTLR (Yu et al., 2011; Fotso, 2018) against which any clustering-based method must be evaluated, and (ii) it allows us to compare performance across our three proposed MoE architectures, where the key difference between them is the expressivity of the MoE head.

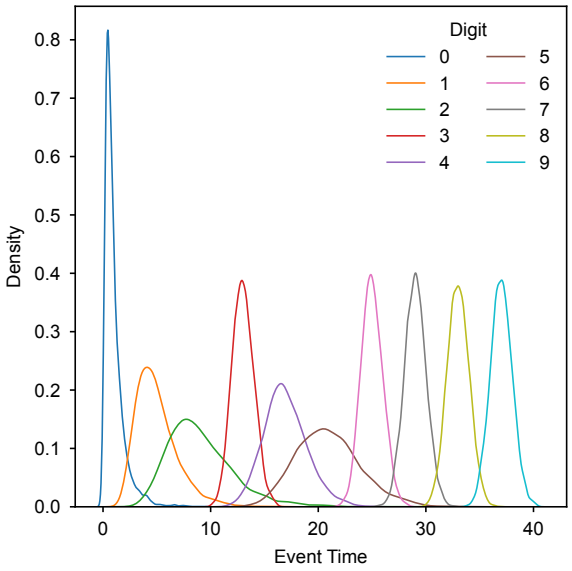


Figure 2: Event distributions for each digit in the Survival MNIST dataset. Each digit has a distinct event distribution, which allows us to evaluate the ability of our models to recover latent groups.

Hyperparameter Sensitivity The second set of experiments varies the number of experts in each MoE architecture to understand how the expressivity of the MoE head impacts the model's performance within each architecture. It also reveals any sensitivity to model specification (i.e., the number of latent

Table 1: We report averages over 5 random seeds and in parentheses average deviation, given a random seed, from the MTLR baseline. Best results per dataset are **bolded**. ↓ indicates lower is better and ↑ indicates higher is better.

Dataset	Model	ECE ↓	Concordance ↑	Brier (25th) ↓	Brier (50th) ↓	Brier (75th) ↓
Survival MNIST	CoxPH	0.030 (0.024)	79.16 (-13.65)	0.112 (0.083)	0.159 (0.125)	0.069 (0.059)
	RSF	0.057 (0.051)	90.06 (-2.76)	0.048 (0.019)	0.073 (0.038)	0.025 (0.015)
	MTLR	0.006 (0.000)	92.82 (0.00)	**0.029 (0.000)**	**0.034 (0.000)**	**0.010 (0.000)**
	Fixed MoE (ours)	0.008 (0.002)	**93.46 (0.65)**	**0.029 (0.000)**	**0.034 (0.000)**	**0.010 (-0.001)**
	Adjustable MoE (ours)	0.008 (0.002)	92.55 (-0.26)	0.030 (0.001)	0.037 (0.003)	0.011 (0.001)
	Personalized MoE (ours)	**0.005 (-0.001)**	92.61 (-0.21)	**0.029 (0.000)**	0.036 (0.002)	**0.010 (0.000)**
SUPPORT2	CoxPH	0.187 (0.130)	78.89 (-1.03)	0.212 (0.055)	0.209 (0.060)	0.236 (0.088)
	RSF	0.187 (0.129)	79.76 (-0.15)	0.207 (0.051)	0.203 (0.055)	0.232 (0.085)
	MTLR	0.057 (0.000)	79.91 (0.00)	0.156 (0.000)	0.149 (0.000)	0.148 (0.000)
	Fixed MoE (ours)	0.054 (-0.004)	79.78 (-0.13)	0.158 (0.001)	0.147 (-0.002)	0.145 (-0.003)
	Adjustable MoE (ours)	**0.048 (-0.009)**	79.83 (-0.08)	0.158 (0.002)	0.145 (-0.003)	0.143 (-0.005)
	Personalized MoE (ours)	**0.048 (-0.009)**	**80.84 (0.93)**	**0.154 (-0.002)**	**0.142 (-0.007)**	**0.138 (-0.009)**
Sepsis	CoxPH	0.635 (0.618)	73.36 (-15.00)	0.272 (0.253)	0.541 (0.508)	0.766 (0.727)
	RSF	0.604 (0.587)	82.69 (-5.67)	0.248 (0.230)	0.603 (0.570)	0.811 (0.773)
	MTLR	0.017 (0.000)	88.36 (0.00)	0.019 (0.000)	0.033 (0.000)	0.039 (0.000)
	Fixed MoE (ours)	0.011 (-0.006)	87.09 (-1.27)	0.019 (0.000)	0.033 (-0.000)	0.039 (0.001)
	Adjustable MoE (ours)	0.009 (-0.008)	88.99 (0.63)	**0.017 (-0.001)**	0.032 (-0.001)	0.037 (-0.002)
	Personalized MoE (ours)	**0.005 (-0.012)**	**89.77 (1.41)**	**0.017 (-0.002)**	**0.030 (-0.003)**	**0.036 (-0.003)**

patient groups). We vary the number of experts from 2 to 20 while holding all other hyperparameters constant.

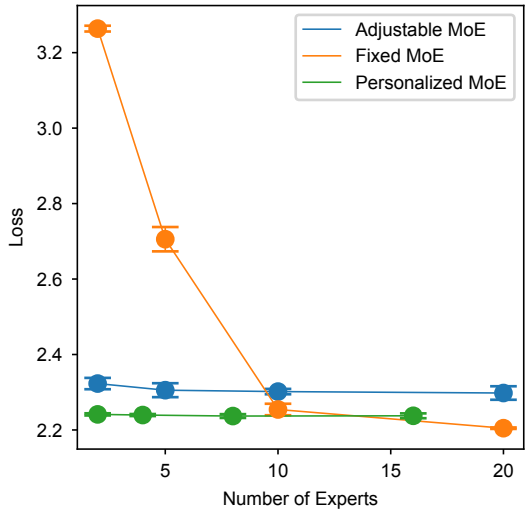
Learned Patient Groups We analyze the routing behavior of the MoE models to understand how patients are assigned to experts. Because Survival MNIST has well-defined latent groups, we can quantify the extent to which each expert specializes in a particular digit. Next, we assess the clinical relevance of patient clustering in a SUPPORT2 Personalized MoE model and quantify the stability of the routing behavior across random seeds using the Adjusted Rand Index (ARI) (Rand, 1971; Hubert and Arabie, 1985).

5. Results

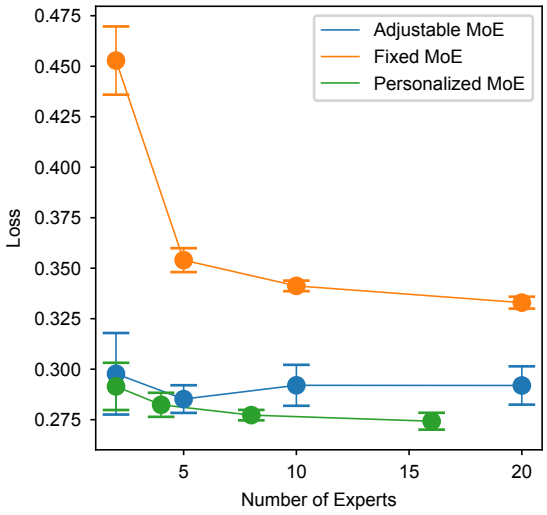
Performance Assessment We report our results in Table 1. The Survival MNIST dataset represents the Platonic ideal of a dataset with clear latent groups and as a result the Fixed MoE model performs best across all metrics except for calibration. This is a setting where the Fixed MoE is perfectly specified, with exactly 10 expert heads, one for each digit. The only information needed to make Bayes-optimal predictions is to identify the digit, at which point the appropriate expert can predict the group's event distribution. Any attempt to adjust or customize the expert distributions per patient only adds unnecessary complexity and can hurt performance. Nevertheless, the Personalized MoE model is best calibrated and matches the performance of the Fixed MoE on Brier at the 25th and 75th percentiles. The Survival MNIST dataset provides an interesting contrast to real-world datasets, where latent groups are almost never as well-defined.

Results on SUPPORT2 show the Personalized MoE model outperforms all other methods across all metrics. Importantly, the Personalized MoE model is outperforming the MTLR model on calibration error, concordance, and Brier score at all time points, while the fixed and adjustable MoE models are, in general, not. On SUPPORT2, our concordance and IPCW Brier scores are in line with prior reports for modern deep survival models (e.g., DSM (Nagpal et al., 2021a)). We see similarly excellent results on the Sepsis dataset, providing further evidence that the Personalized MoE model is able to deliver on all desiderata: clustering, calibration, and predictive accuracy. Consistent with prior work (Nagpal et al., 2021b; Manduchi et al., 2022), we observe that per-patient adjustments can benefit calibration and predictive accuracy as measured by the Brier score relative to MTLR. Our Personalized MoE model extends those gains, delivering simultaneous improvements in calibration, accuracy, and discrimination, achieved with simple end-to-end learning rather than more complex training pipelines (e.g., spline estima-

tion of baseline hazard rates (Nagpal et al., 2021b), EM-based cluster estimation (Buginga and e Silva, 2024), etc.). The Personalized MoE model's ability to form custom event distributions per patient appears to be key to its strong performance on real-world datasets, where latent groups are not as well-defined as in Survival MNIST.



(a) Survival MNIST



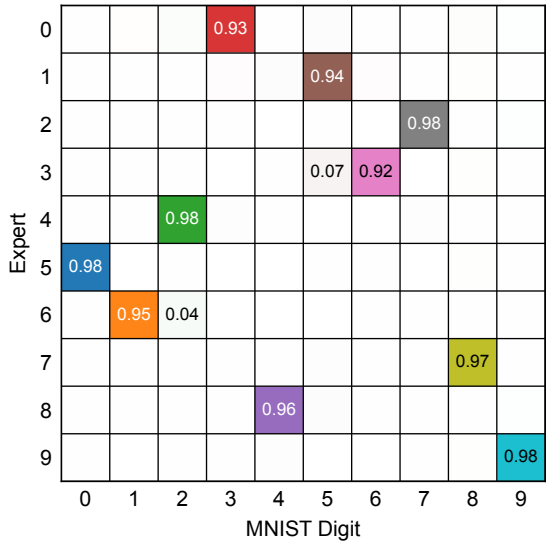
(b) Sepsis

Figure 3: Expert sensitivity analysis over 5 random seeds varying the number of experts. Test set loss as a function of the number of experts.

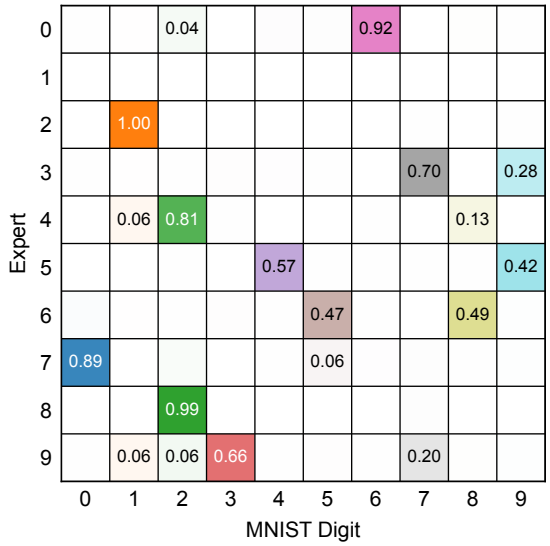
Hyperparameter Sensitivity We report the results of our expert sensitivity analysis in Figure 3. We observe that the Fixed MoE model is highly sensitive to the number of experts used before enough experts are present to capture the latent groups in the data. The adjustable MoE model is less sensitive to the number of experts, as it can adjust the expert distributions per patient even if there are insufficient number of experts. The Personalized MoE model is the least sensitive to the number of experts, as it can form custom event distributions for each patient regardless of the number of experts. We conclude that there is a continuum of sensitivity to model specification based on the expressivity of the experts.

Learned Patient Groups We report the first part of our routing analysis results in Figure 4, where each row of the matrices represents the distribution over Survival MNIST digits routed to that particular expert. For example, in Figure 4(a), among all data points that are routed to expert 0 in a Fixed MoE, 93% are digit 3. Similarly, for expert 1, 94% are digit 5, and so on. On the Survival MNIST dataset, we observe a strong degree of specialization per expert in the Fixed MoE model. The Personalized MoE model's routing decisions in Figure 4(b) show a slightly lower, but still strong, degree of specialization per expert. For Figure 4, the routing decision for a data instance is made by selecting the expert with the highest weight. This degree of specialization per expert in both models indicates that these models are able to recover latent groups in the Survival MNIST dataset.

We also analyze the clustering behavior of a Personalized MoE model trained on SUPPORT2 in Figure 5 to better understand how patient groups are formed on a real-world dataset. We see a variety of clusters sizes (Figure 5(a)), indicating that Cluster 1 could possibly be merged into another cluster given its small size. We also find that patient risks are separated by cluster as evidenced by Figure 5(b). We then investigate kernel density estimates of continuous features (e.g., age in Figure 5(c)) broken down by cluster. For categorical features, we define a z-score based on the proportion of patients in a cluster that have a given categorical feature value relative to the overall proportion of patients with that feature value (see Appendix Section J for more details). Importantly, these analyses allowed us to define succinct, clinician-informed signatures for each cluster, which are summarized in Table 2 and described in more de-



(*a*) Fixed MoE Clustering



(*b*) Personalized MoE Clustering

Figure 4: MNIST digit clustering by expert in (a) a **Fixed MoE** model and (b) a **Personalized MoE** model.

tail in Appendix Section I. This indicates that the Personalized MoE model is able to discover clinically meaningful patient groups in real-world datasets.

In Appendix Figure 8, we display a detailed breakdown of two patients' routing behavior and predictions on all three MoE models, which provides some evidence that the routing behavior of the Fixed MoE and Adjustable MoE models is event time based, which is consistent with how these models are constructed.

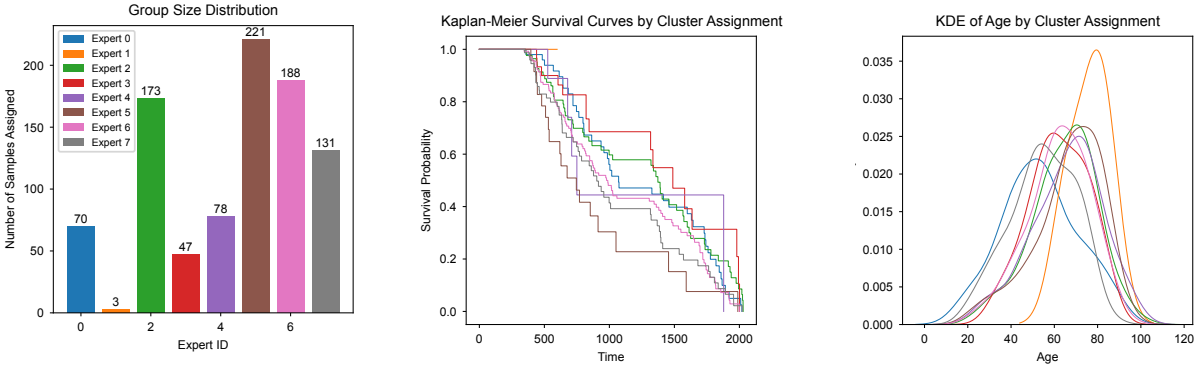
Table 2: Cluster sizes, Kaplan-Meier (KM) risk tier, and succinct signatures (clinician-informed). Vitals and labs are from day 3 of the hospital stay.

Cluster (Size)	KM risk tier	Signature
C_0 ($n = 70$)	mid	50s; ARF/MOSF ± sepsis; low income (2nd lowest)
C_1 ($n = 3$)	n/a (tiny)	80s; metastatic lung cancer; anecdotal size
C_2 ($n = 173$)	mid	70s; older White; lung CA/COPD/CHF/cirrhosis
C_3 ($n = 47$)	**lowest**	50-60s men; metastatic colon CA; ESRD/HD with low ADLs; best vitals/pH
C_4 ($n = 78$)	varied	70s; older Black; ARF/MOSF ± sepsis; DNR (odd KM drop)
C_5 ($n = 221$)	**highest**	80s; dementia; *early* DNR; ARF/MOSF ± sepsis/coma
C_6 ($n = 188$)	mid	60s; COPD/CHF/cirrhosis; diabetes; *lowest* income
C_7 ($n = 131$)	elevated	50s; ARF/MOSF ± sepsis; coma/intubated; low income; 2-mo risk undercalled

We also measure the stability of the routing behavior for the Personalized MoE models across random seeds. We define stability as the degree to which patients are routed to the same expert across different random seeds. Using the Top-1 routing rule (the most active expert for the patient defines its cluster assignment), we find a moderate degree of stability as measured by the Adjusted Rand Index (ARI) with a value of 0.36, indicating that the Personalized MoE model is able to route patients to the same expert across different random seeds moderately consistently.

6. Conclusion

We introduce three discrete-time deep mixture-of-experts (MoE) based survival architectures, one of which achieves excellent clustering, calibration error, and Brier scores at key points in time. We have shown



(*a*) Group sizes (*b*) Survival curves (*c*) Age by cluster

Figure 5: Routing of 911 unseen patients in the SUPPORT2 dataset to experts in a Personalized MoE model. (a) shows the group sizes (by Top-1 expert activation), (b) shows the survival curves for patients routed to each expert, and (c) shows the age distribution of patients in each cluster. The Personalized MoE model is able to discover clinically meaningful patient groups in real-world datasets by risk-profile and patient attributes.

that the expressiveness of the experts is a key differentiator between deep MoE-based survival analysis models, supported by a targeted set of experiments on these three architectures. We observe that the Personalized MoE model is able to deliver on all desiderata: clustering, calibration, and predictive accuracy as evidenced by results on three survival analysis datasets, including two real-world datasets. We explore sensitivity to the number of experts and find that the Personalized MoE model is the least sensitive to the number of experts, making it a robust choice for survival analysis tasks. We also analyze the routing behavior of the MoE models and use a synthetic survival analysis dataset to validate that the models are able to recover latent groups in the data. In practice, we find that the Personalized MoE model is able to discover clinically meaningful patient groups in real-world datasets. Overall, the Personalized MoE model provides a powerful and flexible approach to survival analysis that can adapt to the complexities of real-world clinical data and patient heterogeneity. It maintains strong performance across multiple metrics of interest to clinicians and researchers alike, including calibration and predictive accuracy, and enables reasoning by analogy to similar patients.

7. Limitations

This study does not claim state-of-the-art across all survival benchmarks or architectures. Our claims are relative and internal: within a controlled setting, increasing expert expressivity improves calibration and accuracy on two real-world datasets, with a synthetic counterexample (Survival MNIST) illustrating when fixed prototypes are optimal. Results should be interpreted in this scope; adding further model classes (e.g., DeepHit, continuous-time parametric mixtures, etc.) is orthogonal to our central question and left for future comparative work.

Acknowledgments

We would like to thank Jake Snell, Thomas Zollo, Zhun Deng, Kayla Schiffer-Kane, and Emily Saunders for their helpful discussions. This publication was supported by the National Center for Advancing Translational Sciences, National Institutes of Health, through Grant Number UL1TR001873. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. The authors also acknowledge support from the National Science Foundation and by DoD OUSD (R&E) under Cooperative Agreement PHY-2229929 (The NSF AI Institute for Artificial and Natural Intelligence) (RZ). This work was also supported in part

by National Institutes of Health: 1R01NS129760-01 (SP), 1R01NS131606-01 (SP), American Heart Association: 24SCEFIA1259295 (MM).

References

Md Manjurul Ahsan, Shahana Akter Luna, and Zahed Siddique. Machine-Learning-Based Disease Diagnosis: A Comprehensive Review. *Healthcare*, 10(3):541, March 2022. ISSN 2227-9032. doi: 10.3390/healthcare10030541.

Ahmed M Alaa and Mihaela van der Schaar. Deep multi-task gaussian processes for survival analysis with competing risks. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 2326–2334, 2017.

Gabriel Buginga and Edmundo de Souza e Silva. Clustering Survival Data using a Mixture of Nonparametric Experts, May 2024.

Gabriele Campanella, Lucas Kook, Ida Häggström, Torsten Hothorn, and Thomas J. Fuchs. Deep conditional transformation models for survival analysis, October 2022.

Alfred F. Connors, Jr, Neal V. Dawson, Norman A. Desbiens, William J. Fulkerson, Jr, Lee Goldman, William A. Knaus, Joanne Lynn, Robert K. Oye, Marilyn Bergner, Anne Damiano, Rosemarie Hakim, Donald J. Murphy, Joan Teno, Beth Virnig, Douglas P. Wagner, Albert W. Wu, Yutaka Yasui, Detra K. Robinson, Barbara Kreling, Jennie Dulac, Rose Baker, Sam Holayel, Thomas Meeks, Mazen Mustafa, Juan Vegarra, Carlos Alzola, Frank E. Harrell, Jr, E. Francis Cook, Mary Beth Hamel, Lynn Peterson, Russell S. Phillips, Joel Tsevat, Lachlan Forrow, Linda Lesky, Roger Davis, Nancy Kressin, Jeanmarie Solzan, Ann Louise Puopolo, Laura Quimby Barrett, Nora Bucko, Deborah Brown, Maureen Burns, Cathy Foskett, Amy Hozid, Carol Keohane, Colleen Martinez, Dorcie McWeeney, Debra Melia, Shelley Otto, Kathy Sheehan, Alice Smith, Lauren Tofias, Bernice Arthur, Carol Collins, Mary Cunnion, Deborah Dyer, Corinne Kulak, Mary Michaels, Maureen O'Keefe, Marian Parker, Lauren Tuchin, Dolly Wax, Diana Weld, Liz Hiltunen, Georgie Marks, Nancy Mazzapica, Cindy Medich, Jane Soukup, Robert M. Califf, Anthony N. Galanos, Peter Kussin, Lawrence H. Muhlbaier, Maria Winchell, Lee Mallatratt, Ella Akin, Lynne Belcher, Elizabeth Buller, Eileen Clair, Laura Drew, Libby Fogelman, Dianna Frye, Beth Fraulo, Debbie Gessner, Jill Hamilton, Kendra Kruse, Dawn Landis, Louise Nobles, Rene Oliverio, Carroll Wheeler,

Nancy Banks, Steven Berry, Monie Clayton, Patricia Hartwell, Nan Hubbard, Isabel Kussin, Barbara Norman, Jackie Noveau, Heather Read, Barbara Warren, Jane Castle, Kathy Turner, Rosalie Perdue, Claudia Coulton, C. Seth Landefeld, Theodore Speroff, Stuart Youngner, Mary J. Kennard, Mary Naccaratto, Mary Jo Roach, Maria Blinkhorn, Cathy Corrigan, Elsie Geric, Laura Haas, Jennifer Ham, Julie Jerdonek, Marilyn Landy, Elaine Marino, Patti Olesen, Sherry Patzke, Linda Repas, Kathy Schneeberger, Carolyn Smith, Colleen Tyler, Mary Zenczak, Helen Anderson, Pat Carolin, Cindy Johnson, Pat Leonard, Judy Leuenberger, Linda Palotta, Millie Warren, Jane Finley, Toni Ross, Gillian Solem, Sue Zronek, Sara Davis, Steven Broste, Peter Layde, Michael Kryda, Douglas J. Reding, Humberto J. Vidaillet, Jr, Marilyn Folien, Patsy Mowery, Barbara E. Backus, Debra L. Kempf, Jill M. Kupfer, Karen E. Maassen, Jean M. Rohde, Nancy L. Wilke, Sharon M. Wilke, Elizabeth A. Albee, Barbara Backus, Angela M. Franz, Diana L. Henseler, Juanita A. Herr, Irene Leick, Carol L. Lezotte, Laura Meddaugh, Linda Duffy, Debrah Johnson, Susan Kronenwetter, Anne Merkel, Paul E. Bellamy, Jonathan Hiatt, Neil S. Wenger, Margaret Leal-Sotelo, Darice Moranville-Hawkins, Patricia Sheehan, Diane Watanabe, Myrtle C. Yamamoto, Allison Adema, Ellen Adkins, Ann-Marie Beckson, Mona Carter, Ellen Duerr, Ayam El-Hadad, Ann Farber, Ann Jackson, John Justice, Agnes O'Meara, Lee Benson, Lynette Cheney, Carlo Medina, Jane Moriarty, Kay Baker, Cleine Marsden, Kara Watne, Diane Goya, Norman Desbiens, William J. Fulkerson, Charles C. J. Carpenter, Ronald A. Carson, Don E. Detmer, Donald E. Steinwachs, Vincent Mor, Robert A. Harootyan, Alex Leaf, Rosalyn Watts, Sankey Williams, and David Ransohoff. A Controlled Trial to Improve Care for Seriously III Hospitalized Patients: The Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatments (SUPPORT). *JAMA*, 274(20):1591–1598, November 1995. ISSN 0098-7484. doi: 10.1001/jama.1995.03530200027032.

D. R. Cox. Regression Models and Life-Tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972. ISSN 2517-6161. doi: 10.1111/j.2517-6161.1972.tb00899.x.

Fatma A. Eltawil, Michael Atalla, Emily Boulos, Afsaneh Amirabadi, and Pascal N. Tyrrell. Analyzing Barriers and Enablers for the Acceptance of Artificial Intelligence Innovations into Radiology Practice: A Scoping Review. *Tomography*, 9(4): 1443–1455, August 2023. ISSN 2379-139X. doi: 10.3390/tomography9040115.

Bradley J. Erickson, Panagiotis Korfiatis, Zeynettin Akkus, and Timothy L. Kline. Machine Learning for Medical Imaging. *RadioGraphics*, 37(2):505–515, March 2017. ISSN 0271-5333. doi: 10.1148/rg.2017160130.

Stephane Fotso. Deep Neural Networks for Survival Analysis Based on a Multi-Task Framework, January 2018.

Thomas A. Gerds and Martin Schumacher. Consistent estimation of the expected Brier score in general survival models with right-censored event times. *Biometrical Journal. Biometrische Zeitschrift*, 48(6):1029–1040, December 2006. ISSN 0323-3847. doi: 10.1002/bimj.200610301.

Humza Haider, Bret Hoehn, Sarah Davis, and Russell Greiner. Effective ways to build and evaluate individual survival distributions. *J. Mach. Learn. Res.*, 21(1):85:3289–85:3351, January 2020. ISSN 1532-4435.

F. E. Harrell, K. L. Lee, and D. B. Mark. Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, 15(4):361–387, February 1996. ISSN 0277-6715. doi: 10.1002/(SICI)1097-0258(19960229)15:4⟨361::AID-SIM168⟩3.0.CO;2-4.

Katharine E. Henry, David N. Hager, Tiffany M. Osborn, Albert W. Wu, and Suchi Saria. Comparison of Automated Sepsis Identification Methods and Electronic Health Record–based Sepsis Phenotyping: Improving Case Identification Accuracy by Accounting for Confounding Comorbid Conditions. *Critical Care Explorations*, 1(10):e0053, October 2019. ISSN 2639-8028. doi: 10.1097/CCE.0000000000000053.

Torsten Hothorn, Lisa Möst, and Peter Bühlmann. Most Likely Transformations. *Scandinavian Journal of Statistics*, 45(1):110–134, March 2018. ISSN 0303-6898, 1467-9469. doi: 10.1111/sjos.12291.

Bojian Hou, Hongming Li, Zhicheng Jiao, Zhen Zhou, Hao Zheng, and Yong Fan. Deep Clustering Survival Machines with Interpretable Expert Distributions. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, pages 1–4, April 2023. doi: 10.1109/ISBI53787.2023.10230844.

Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of Classification*, 2(1):193–218, December 1985. ISSN 1432-1343. doi: 10.1007/BF01908075.

Hemant Ishwaran, Udaya B. Kogalur, Eugene H. Blackstone, and Michael S. Lauer. Random survival forests. *The Annals of Applied Statistics*, 2(3): 841–860, September 2008. ISSN 1932-6157, 1941-7330. doi: 10.1214/08-AOAS169.

Sadiya S. Khan, Kunihiro Matsushita, Yingying Sang, Shoshana H. Ballew, Morgan E. Grams, Aditya Surapaneni, Michael J. Blaha, April P. Carson, Alexander R. Chang, Elizabeth Ciemins, Alan S. Go, Orlando M. Gutierrez, Shih-Jen Hwang, Simerjot K. Jassal, Csaba P. Kovesdy, Donald M. Lloyd-Jones, Michael G. Shlipak, Latha P. Palaniappan, Laurence Sperling, Salim S. Virani, Katherine Tuttle, Ian J. Neeland, Sheryl L. Chow, Janani Rangaswami, Michael J. Pencina, Chiadi E. Ndumele, Josef Coresh, and for the Chronic Kidney Disease Prognosis Consortium and the American Heart Association Cardiovascular-Kidney-Metabolic Science Advisory Group. Development and Validation of the American Heart Association's PREVENT Equations. *Circulation*, 149(6):430–449, February 2024. doi: 10.1161/CIRCULATIONAHA.123.067626.

Håvard Kvamme and Ørnulf Borgan. Continuous and discrete-time survival prediction with neural networks. *Lifetime Data Analysis*, 27(4):710–736, October 2021. ISSN 1572-9249. doi: 10.1007/s10985-021-09532-6.

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 AAAI Conference on Artificial Intelligence*, 32(1), April 2018. ISSN 2374-3468. doi: 10.1609/aaai.v32i1.11842.

Laura Manduchi, Ričards Marcinkevičs, Michela C. Massi, Thomas Weikert, Alexander Sauter, Verena Gotta, Timothy Müller, Flavio Vasella, Marian C. Neidert, Marc Pfister, Bram Stieltjes, and Julia E. Vogt. A Deep Variational Approach to Clustering Survival Data. In *International Conference on Learning Representations*, March 2022. doi: 10.48550/arXiv.2106.05763.

Xenia Miscouridou, Adler Perotte, Noémie Elhadad, and Rajesh Ranganath. Deep survival analysis: Nonparametrics and missingness. In *Machine Learning for Healthcare Conference*, pages 244–256. PMLR, 2018.

Mélodie Monod, Peter Krusche, Qian Cao, Berkman Sahiner, Nicholas Petrick, David Ohlssen, and Thibaud Coroller. TorchSurv: A Lightweight Package for Deep Survival Analysis. *Journal of Open Source Software*, 9(104):7341, December 2024. ISSN 2475-9066. doi: 10.21105/joss.07341.

Chirag Nagpal, Xinyu Li, and Artur Dubrawski. Deep Survival Machines: Fully Parametric Survival Regression and Representation Learning for Censored Data With Competing Risks. *IEEE Journal of Biomedical and Health Informatics*, 25(8): 3163–3175, August 2021a. ISSN 2168-2208. doi: 10.1109/JBHI.2021.3052441.

Chirag Nagpal, Steve Yadlowsky, Negar Rostamzadeh, and Katherine Heller. Deep Cox Mixtures for Survival Regression. In *Proceedings of the 6th Machine Learning for Healthcare Conference*, pages 674–708. PMLR, October 2021b.

Patrick A. Norman, Wanlu Li, Wenyu Jiang, and Bingshu E. Chen. deepAFT: A nonlinear accelerated failure time model with artificial neural network. *Statistics in Medicine*, 43(19):3689–3701, 2024. ISSN 1097-0258. doi: 10.1002/sim.10152.

Sebastian Pölsterl. Scikit-survival: A Library for Time-to-Event Analysis Built on Top of scikit-learn. *Journal of Machine Learning Research*, 21 (212):1–6, 2020. ISSN 1533-7928.

Shi-ang Qi, Weijie Sun, and Russell Greiner. SurvivalEVAL: A Comprehensive Open-Source Python Package for Evaluating Individual Survival Distributions. *Proceedings of the AAAI Symposium Series*, 2(1):453–457, 2023. ISSN 2994-4317. doi: 10.1609/aaaiss.v2i1.27713.

Shi-Ang Qi, Yakun Yu, and Russell Greiner. Conformalized Survival Distributions: A Generic Post-Process to Increase Calibration. In *Proceedings of the 41st International Conference on Machine Learning*, pages 41303–41339. PMLR, July 2024.

William M. Rand. Objective Criteria for the Evaluation of Clustering Methods. *Journal of the American Statistical Association*, 66(336):846–850, December 1971. ISSN 0162-1459. doi: 10.1080/01621459.1971.10482356.

Rajesh Ranganath, Adler Perotte, Noémie Elhadad, and David Blei. Deep survival analysis. In *Machine Learning for Healthcare Conference*, pages 101–114. PMLR, 2016.

Matthew A. Reyna, Christopher S. Josef, Russell Jeter, Supreeth P. Shashikumar, M. Brandon Westover, Shamim Nemati, Gari D. Clifford, and Ashish Sharma. Early Prediction of Sepsis From Clinical Data: The PhysioNet/Computing in Cardiology Challenge 2019. *Critical Care Medicine*, 48 (2):210–217, February 2020. ISSN 0090-3493. doi: 10.1097/CCM.0000000000004145.

Rebecca Roelofs, Nicholas Cain, Jonathon Shlens, and Michael C. Mozer. Mitigating Bias in Calibration Error Estimation. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, pages 4036–4054. PMLR, May 2022.

Mervyn Singer, Clifford S. Deutschman, Christopher Warren Seymour, Manu Shankar-Hari, Djillali Annane, Michael Bauer, Rinaldo Bellomo, Gordon R. Bernard, Jean-Daniel Chiche, Craig M. Coopersmith, Richard S. Hotchkiss, Mitchell M. Levy, John C. Marshall, Greg S. Martin, Steven M. Opal, Gordon D. Rubenfeld, Tom van der Poll, Jean-Louis Vincent, and Derek C. Angus. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *JAMA*, 315(8): 801–810, February 2016. ISSN 0098-7484. doi: 10.1001/jama.2016.0287.

Chun-Nam Yu, Russell Greiner, Hsiu-Chin Lin, and Vickie Baracos. Learning Patient-Specific Cancer Survival Distributions as a Sequence of Dependent Regressors. In *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011.

Qixian Zhong, Jonas W Mueller, and Jane-Ling Wang. Deep Extended Hazard Models for Survival Analysis. In *Advances in Neural Information Processing Systems*, volume 34, pages 15111–15124. Curran Associates, Inc., 2021.

Appendix A. Additional Discussion & Limitations

There are limits to the number of experts that can be used in a Personalized MoE model for a given hidden dimension because: (i) the hidden dimension must be divisible by the number of experts, and (ii) the hidden dimension must be large enough to represent all discrete event times. Event times may cluster at key points in time, which would lower the dimensionality of the representation required to correctly decode. Our experiments work with 100 discrete event times and typically use a representation of size 16 (128 hidden dimensions / 8 experts), which is sufficient to represent the event time distribution for the datasets we explore, perhaps as a result of expert specialization by event time. If the event time distribution is more complex, then one may need number of dimensions on the order of number of time bins to form a sufficiently expressive event distribution. Another limitation is that we do not explore time-varying inputs and outputs (i.e., multiple predictions per patient as their input features are updated). In principle, all of our methods can be attached to a recurrent neural network or Transformer but the details of how to interpret groups of patients in a time-varying setting is an open question. Another pro/con of the Personalized MoE is that it is not as sensitive to the number of experts as the other two models, which is limiting if the primary goal is to discover latent groups. One heuristic is to use the Fixed MoE to discover the number of latent groups (e.g., using the elbow method shown in Figure 3) and then use that number of experts in the Personalized MoE model. Another approach is to use a large number of experts in the Personalized MoE model and then prune/consolidate experts based on their routing behavior, but we leave this for future work. In future work, we are also interested in exploring simpler (yet expressive) transformation families for the Adjustable MoE that are monotone and naturally map from $[0, 1]$ to $[0, 1]$.

Appendix B. Loss Functions

B.1. Uncensored Loss

All of our models are optimized using the Multitask Logistic Regression Loss (MTLR) (Yu et al., 2011; Fotso, 2018). We encode event times such that if patient i has a medical event at time s then all times $t \geq s$ will have label $y_t^i = 1$. Correspondingly, times $t < s$ will be labeled with 0. Typical label strings will look like a sequence of 0s followed by a sequence of 1s (e.g., $(0, 0, 0, 1, 1)$). The probability that our model assigns to the ground truth label sequence for a particular patient (the likelihood function) is then given as

$$p(\mathbf{Y} = (y_1, y_2, \ldots, y_m)|\mathbf{z}) = \frac{\exp(\sum_{j=1}^{m} y_j z_j)}{\sum_{k=0}^{m} \exp(f(\mathbf{z}, k))} \tag{14}$$

where $\mathbf{z} \in \mathbb{R}^m$ are logits from the model and $f(\mathbf{z}, k) = \sum_{j=0}^{k} 0 \cdot z_j + \sum_{j=k+1}^{m} 1 \cdot z_j$ for $0 \leq k \leq m$, which constrains the space of valid event sequences to runs of 0s followed by runs of 1s and corresponds to the disease occurring in the interval $[k, k+1)$. The boundary case is $f(\mathbf{z}, m) = 0$, which corresponds to the sequence of all 0s. We can minimize the negative log-likelihood for an uncensored patient with the following loss

$$\mathcal{L}_{\text{uncensored}} = -\sum_{j=1}^{m} y_j z_j - \log\left(\sum_{k=0}^{m} \exp(f(\mathbf{z}, k))\right) \tag{15}$$

We now provide more detail on how to exchange between the model's logits $\mathbf{z}$ and a PMF $\mathbf{p} \in [0, 1]^m$ over event times. Our model produces *increment logits* due to its interaction with the cumulative-sum parameterization of the softmax (Equation 14). However, all methods must blend distributions over event times from different experts in probability space and therefore we rely on Equation 14 to map from increment logit space to probability space. We must also define an inverse operation to map from a PMF back to increment logits. We have

$$z_j = \log(p_j) - \log(p_{j+1}) \text{ for } j = 1, \ldots, m-1 \tag{16}$$

and set $z_m = 0$. We now derive this inverse operation for completeness. Let

$$u_t = \sum_{j=t}^{m} z_j \text{ for } t = 1, \ldots, m. \tag{17}$$

By Equation 14, $p_j = \frac{\exp(u_j)}{\sum_{k=0}^{m} \exp(f(\mathbf{z},k))}$ for $j = 1, \ldots, m$, which is equivalent to $\text{softmax}(\mathbf{u})$ for $\mathbf{u} \in \mathbb{R}^m$. Recall that the softmax function is invariant to additive shifts (i.e., $\text{softmax}(\mathbf{u}) = \text{softmax}(\mathbf{u} + c)$ for any constant c). We can therefore set $c = -u_m$ so that $u_m = 0$ and by Equation 17, $z_m = 0$. We

can now write $u_j = \log(p_j) + c'$ for $j = 1, \ldots, m$. Since $u_m = 0$, we have $c' = -\log(p_m)$ and therefore $u_j = \log(p_j) - \log(p_m)$ for $j = 1, \ldots, m$. By Equation 17, we have

$$z_j = u_j - u_{j+1} \tag{18}$$
$$= \log(p_j) - \log(p_m) - \log(p_{j+1}) + \log(p_m) \tag{19}$$
$$= \log(p_j) - \log(p_{j+1}) \text{ for } j = 1, \ldots, m-1 \tag{20}$$

and $z_m = 0$, which completes the derivation.

B.2. Censored Loss

We now describe how to contend with censored data, which occurs when we do not observe if or when a patient has an event. Suppose a patient is censored at time s_c and time $t + t_j$ is the closest time point after s_c. Then all sequences $\mathbf{Y} = (y_1, y_2, \ldots, y_m)$ with $y_i = 0$ for $i < j$ are consistent with this censored observation. Therefore the likelihood for a censored patient is the survival function (i.e., 1 minus the cumulative density function (CDF)), which is

$$p(S \geq t + t_j | \mathbf{z}) = \frac{\sum_{k=j}^{m} \exp(f(\mathbf{z}, k))}{\sum_{k=0}^{m} \exp(f(\mathbf{z}, k))} \tag{21}$$

and in turn, the negative log-likelihood loss for a censored patient is

$$\mathcal{L}_{\text{censored}} = -\left[\log\left(\sum_{k=j}^{m} \exp(f(\mathbf{z}, k))\right) - \log\left(\sum_{k=0}^{m} \exp(f(\mathbf{z}, k))\right)\right]. \tag{22}$$

B.3. Regularizers

We apply a load-balancing loss to all MoE models to ensure the model uses all available experts. Let $\bar{\boldsymbol{\alpha}} = \frac{1}{b}\sum_{i=1}^{b} \boldsymbol{\alpha}(\mathbf{x}^i)$ be the average expert distribution over a batch of size b. The load-balancing loss encourages the model to use all experts equally across the batch of examples by penalizing low-entropy average expert distributions. For a given batch, the loss is given as

$$\mathcal{L}_{\text{load-balance}} = \lambda_{\text{lb}} \cdot n \sum_{i=1}^{n} \bar{\alpha}_i^2, \tag{23}$$

where n is the number of experts and λ_{lb} is a hyperparameter that controls the strength of the regularization. This loss is minimized when $\bar{\boldsymbol{\alpha}}$ is the uniform distribution.

Appendix C. Hyperparameters and Training Details

Table 3: Parameter counts for all neural methods.

Dataset	Model	Parameters
Survival MNIST	MTLR	187,189
	Fixed MoE	209,844
	Adjustable MoE	194,883
	Personalized MoE	195,891
SUPPORT2	MTLR	68,521
	Fixed MoE	69,480
	Adjustable MoE	69,435
	Personalized MoE	62,141
Sepsis	MTLR	62,945
	Fixed MoE	63,008
	Adjustable MoE	63,579
	Personalized MoE	57,909

Table 4: Model specific hyperparameters for Survival MNIST.

Model	Fixed MoE	Adjustable MoE	Personalized MoE	MTLR
Hidden Dim. (h)	208	186	160	176
Num. Layers (ℓ)	2	2	1	2
Num. Experts (n)	10	10	10	-
Learning Rate	5e-4	5e-4	5e-4	5e-4

Table 5: Model specific hyperparameters for SUPPORT2 datasets.

Model	Fixed MoE	Adjustable MoE	Personalized MoE	MTLR
Hidden Dim. (h)	176	186	128	176
Num. Layers (ℓ)	2	2	1	2
Num. Experts (n)	10	10	8	-
Learning Rate	5e-3	5e-3	5e-4	5e-4

We define a validation set for Survival MNIST by randomly sampling 5,000 examples from the training set and then use the provided test set for final evaluation. For SUPPORT2 and Sepsis, we randomly sample 10% of all examples to form a validation set and sample another 10% to form a test set. We use the validation set loss to perform hyperparameter tuning for each dataset and method. The temperature parameter κ is initialized to 2.0 for all MoE models and learned during training. The load balancing

Table 6: Model specific hyperparameters for Sepsis datasets.

Model	Fixed MoE	Adjustable MoE	Personalized MoE	MTLR
Hidden Dim. (h)	176	186	128	176
Num. Layers (ℓ)	2	2	1	2
Num. Experts (n)	10	10	8	-
Learning Rate	5e-4	5e-4	5e-4	5e-4

Table 7: Shared hyperparameters for all models.

Hyperparameter	Value
Batch Size	64
λ_{lb}	0.01
κ Init.	2.0
Time Bins (m)	100

Table 8: Random survival forest hyperparameters using the `scikit-survival` library.

Hyperparameter	Survival MNIST	Sepsis	SUPPORT2
`n_estimators`	100	50	200
`max_features`	`sqrt`	`sqrt`	`sqrt`
`min_samples_split`	100	50	200

Table 9: Cox proportional hazards hyperparameters for all datasets using the `scikit-survival` library.

Hyperparameter	Value
`fit_baseline`	`True`
`alphas`	`[0.01]`
`l1_ratio`	`0.01`

loss λ_{lb} is constrained. For instance, setting the load balancing loss to be 0 or too close to 0 results in the model not using all of the experts. We found $\lambda_{\text{lb}} = 0.01$ to consistently allow models to use most, if not all, the experts, while still allowing for specialization. We tuned the learning rate for all neural methods from the set $\{5\text{e-}3, 5\text{e-}4, 5\text{e-}5\}$. All models are trained with the Adam optimizer. For the Cox Proportional Hazards model we tuned the `alphas` hyperparameter from the set $\{0.001, 0.01, 0.1\}$ and the `l1_ratio` from the set $\{0.01, 0.5, 1.0\}$. The Random Survival Forest model's `n_estimators` hyperparameter was tuned from the set $\{50, 100, 200\}$ and the `min_samples_split` hyperparameter was tuned from the set $\{50, 100, 200\}$. Due to long runtimes and high memory requirements on the Survival MNIST dataset–likely due to the large number of features–we directly set `n_estimators` and `min_samples_split` to 100 and 100 respectively without tuning. We set hidden dimension sizes so that parameter counts are approximately equal across neural methods to ensure a fair comparison and isolate the effects of the decoder head used. We use a hidden dimension of 128-208 and 1-2 hidden layers for all models, which are fully connected layers followed by ReLU activations. We use 8-10 experts for all MoE models. The number of discrete time bins is set to $m = 100$ for all datasets. We train all models to convergence as measured by the validation set loss, using early stopping with a patience of 10 epochs. All models are implemented in PyTorch and trained on a single GPU. Our Github code repository is available at https://github.com/ToddMorrill/survival-moe. We report the final hyperparameters used for each model and dataset in the tables above.

Appendix D. Survival MNIST Data Generation

Setup Let $\{(x_i, y_i)\}_{i=1}^{N}$ be MNIST images x_i with class labels $y_i \in \{0, \ldots, 9\}$. For each class k, we fix a target *event-time* mean $m_k > 0$ and standard deviation $s_k > 0$ for a log-normal model of the event time $T \mid Y = k$. In our experiments we use:

$$\boldsymbol{m} = (1, 5, 9, 13, 17, 21, 25, 29, 33, 37),$$
$$\boldsymbol{s} = (1, 2, 3, 1, 2, 3, 1, 1, 1, 1).$$

Event-time model For class k, we draw a *raw* event time

$$T^\star \mid (Y = k) \sim \text{LogNormal}(\mu_k, \sigma_k^2), \quad \text{with pdf}$$
$$f_{T|k}(t) = \frac{1}{t\,\sigma_k\sqrt{2\pi}} \exp\left(-\frac{(\ln t - \mu_k)^2}{2\sigma_k^2}\right), \quad t > 0,$$

and CDF $F_{T|k}(t) = \Phi\left(\frac{\ln t - \mu_k}{\sigma_k}\right)$. We choose (μ_k, σ_k) so that the resulting log-normal has *mean* m_k and *standard deviation* s_k on the original time scale. Using $\mathbb{E}[T] = e^{\mu+\sigma^2/2}$ and $\text{Var}(T) = (e^{\sigma^2} - 1)e^{2\mu+\sigma^2}$, the code computes

$$\mu_k = \log\left(\frac{m_k^2}{\sqrt{s_k^2 + m_k^2}}\right), \qquad \sigma_k = \sqrt{\log\left(1 + \frac{s_k^2}{m_k^2}\right)}.$$

(Equivalently, SciPy's `lognorm` is called with shape $s = \sigma_k$, scale $= \exp(\mu_k)$, and `loc`$= 0$.)

Right-censoring We impose random right-censoring at rate $p_c = 0.15$. Let $C \subset \{1, \ldots, N\}$ be a uniformly random subset of size $\lfloor p_c N \rfloor$. For $i \notin C$ (uncensored), we set the observed time $t_i = T_i^\star$ and event indicator $\delta_i = 1$. For $i \in C$ (censored), we draw a censoring time

$$c_i \sim \text{Uniform}(0,\, T_i^\star)$$

and set $t_i = c_i$, $\delta_i = 0$. Thus the observed triplet is (t_i, δ_i, y_i) with $t_i > 0$ and $\delta_i \in \{0, 1\}$.

Appendix E. Evaluation Metrics

E.1. Equal-Mass Expected Calibration Error (ECE) for Discrete-Time Survival

Notation We evaluate calibration over a discrete grid $\{t_1, \ldots, t_T\}$ (here $T = 100$). For a batch of N individuals, the model outputs *increment logits* which are mapped to a discrete-time CDF via a monotone transformation:

$$\widehat{F}_{ij} \;\equiv\; \widehat{F}(t_j \mid x_i) \in [0, 1],$$

where $\widehat{F}_{ij}$ follows from Equation 14. The labels are $y_{ij} \in \{0, 1\}$ with the *monotone* convention (zeros up to the event time, then ones thereafter). Let $c_i \in \{0, 1\}$ denote censoring ($c_i = 1$ means censored) and $\delta_i = 1 - c_i$ the usual event indicator. Let T_i denote the event time.

Equal-mass binning (per time t_j) Fix a number of calibration bins Q (default $Q = 10$). For each t_j:

1. Sort individuals by $\widehat{F}_{ij}$ in ascending order; denote the resulting permutation by σ_j.
2. Let $b = \lfloor N/Q \rfloor$ and $r = N \bmod Q$. Define bin sizes $s_q = b + 1$ for $q \in \{1, \ldots, r\}$ and $s_q = b$ for $q \in \{r + 1, \ldots, Q\}$.
3. Assign the first s_1 sorted individuals to bin $q = 1$, the next s_2 to $q = 2$, etc. Define the one-hot assignment $B_{ijq} = 1$ if individual i falls in bin q at time t_j, else 0. Thus $n_{jq} = \sum_i B_{ijq} = s_q$ and $\sum_q n_{jq} = N$.

Per-bin predicted probability Within each time t_j and bin q,

$$\overline{F}_{jq} \;=\; \frac{\sum_{i=1}^N B_{ijq}\, \widehat{F}_{ij}}{\sum_{i=1}^N B_{ijq}} \;=\; \frac{1}{n_{jq}} \sum_{i=1}^N B_{ijq}\, \widehat{F}_{ij}.$$

Per-bin empirical event probability We estimate the empirical probability that the event has occurred by t_j within bin q, denoted $\overline{Y}_{jq}$.

IPCW adjustment Let $G(\cdot)$ be the censoring survival, estimated by a Kaplan-Meier fit on the censoring process (using c_i as the censoring "event" indicator). Kaplan-Meier estimates are implemented using `torchsurv` (Monod et al., 2024). To avoid division by zero, evaluation times are clamped to the support of G. Define event and survivor masks at each (i, j):

$$E_{ij} = \mathbb{1}\{T_i \le t_j\}\, \delta_i, \qquad S_{ij} = \mathbb{1}\{T_i > t_j\}.$$

Use inverse-probability-of-censoring weights

$$w_i^{\text{evt}} = \frac{1}{G(\min\{T_i, \tau_G\})}, \qquad w_j^{\text{surv}} = \frac{1}{G(t_j)},$$

where τ_G is the largest time at which G is defined. Then

$$\text{num}_{jq} = \sum_{i=1}^N B_{ijq}\, w_i^{\text{evt}}\, E_{ij},$$

$$\text{den}_{jq} = \text{num}_{jq} + \sum_{i=1}^N B_{ijq}\, w_j^{\text{surv}}\, S_{ij},$$

and

$$\overline{Y}_{jq} = \text{num}_{jq} / \text{den}_{jq},$$

ECE aggregation (per time t_j) The equal-mass expected calibration error at time t_j is the bin-count-weighted ℓ_1 gap between predicted and empirical means:

$$\text{ECE}(t_j) = \sum_{q=1}^Q w_{jq}\, |\overline{F}_{jq} - \overline{Y}_{jq}|,$$

$$w_{jq} = \frac{n_{jq}}{\sum_{q'=1}^Q n_{jq'}} = \frac{n_{jq}}{N}.$$

(Weights are included because the last few bins may differ in size by at most one when N is not divisible by Q.)

Finally, ECE can be averaged over all times t_j to yield a single scalar summary statistic:

$$\text{ECE} = \frac{1}{T} \sum_{j=1}^T \text{ECE}(t_j).$$

Appendix F. IPCW Brier Score for Discrete-Time Survival

Notation The setup and notation is similar to the ECE metric above.

Per-time squared residuals Define the per-entry squared residuals

$$r_{ij} \;=\; (\widehat{F}_{ij} - y_{ij})^2.$$

Inverse-probability-of-censoring (IPCW) weights. For evaluation at time t_j, use the standard IPCW weights

$$w_i(t_j) \;=\; \frac{\mathbb{1}\{T_i \le t_j\}\,\delta_i}{G(\min\{T_i, \tau_G\})} \;+\; \frac{\mathbb{1}\{T_i > t_j\}}{G(t_j)}.$$

In the implementation, arguments to G are clamped to τ_G to avoid undefined values.

Brier score (IPCW) The IPCW Brier score at each time t_j is

$$\mathrm{BS}(t_j) \;=\; \frac{1}{N}\sum_{i=1}^{N} w_i(t_j)\, r_{ij}.$$

Finally, the Brier score can be averaged over all times t_j to yield a single scalar summary statistic:

$$\mathrm{BS} = \frac{1}{T}\sum_{j=1}^{T} \mathrm{BS}(t_j).$$

F.1. Concordance Index

We use the default settings for the concordance index (C-index) as implemented in the `SurvivalEval` library (Qi et al., 2023). Their implementation uses uses linear interpolation of the discrete survival curve to find the median survival time for each patient. This value is then negated to form a risk score for each patient, which is then used to compute the C-index.

Appendix G. IPCW-Adjusted Concordance Index

Harrell's Concordance index (C-index) (Harrell et al., 1996) is widely used to evaluate the discriminative performance of survival models (Ranganath et al., 2016; Khan et al., 2024). At moderate levels of censoring (e.g., below 40%) Harrell's C-index is a good estimator of the model's discriminative performance but at higher levels of censoring, it can be biased, which can be adjusted by using inverse probability of censoring weights (IPCW) (Pölsterl, 2020). In our study, SUPPORT2 has the highest rate of censoring at 32%, so as a check on our results, we report IPCW-adjusted C-index in Table 10.

Table 10: IPCW-adjusted C-index on SUPPORT2. Parentheses show average deviation from MTLR baseline over 5 seeds.

Model	IPCW-adjusted C-index ↑
CoxPH	77.20 (-1.56)
RSF	78.51 (-0.25)
MTLR	78.75 (0.00)
Fixed MoE (ours)	78.60 (-0.16)
Adjustable MoE (ours)	78.27 (-0.49)
Personalized MoE (ours)	**79.67 (0.91)**

Appendix H. Inversion and Gradients for the Two-Logistic Warp

A. Inversion by Bisection

Recall the patient- and expert-specific forward map

$$F_{k,\mathbf{x}}(u) = \sum_{r=1}^{2} w_{k,r}(\mathbf{x})\,\sigma(a_{k,r}(\mathbf{x})[u - c_{k,r}(\mathbf{x})]), \quad (24)$$

$$\sigma(z) = \frac{1}{1+e^{-z}}, \quad (25)$$

and its endpoint-normalized version

$$\tilde{F}_{k,\mathbf{x}}(u) \;=\; \frac{F_{k,\mathbf{x}}(u) - F_{k,\mathbf{x}}(0)}{F_{k,\mathbf{x}}(1) - F_{k,\mathbf{x}}(0)} \in [0,1]. \quad (26)$$

We define

$$\phi_{k,\mathbf{x}}(u) = \tilde{F}_{k,\mathbf{x}}(u) \quad \text{and} \quad \psi_{k,\mathbf{x}} = \phi_{k,\mathbf{x}}^{-1}.$$

Since σ is strictly increasing and $w_{k,r}(\mathbf{x}), a_{k,r}(\mathbf{x}) > 0$ with $0 < c_{k,1}(\mathbf{x}) < c_{k,2}(\mathbf{x}) < 1$, $F_{k,\mathbf{x}}$ (and hence $\tilde{F}_{k,\mathbf{x}}$) is strictly increasing on $[0,1]$, so the inverse exists and is unique.

For a given $(k, \mathbf{x}, t_j)$ we obtain $\tau^\star = \psi_{k,\mathbf{x}}(t_j)$ as the unique solution to

$$g(\tau;\theta) = \tilde{F}_{k,\mathbf{x}}(\tau;\theta) - t_j \;=\; 0,$$
$$\theta = (w_{k,1:2}, a_{k,1:2}, c_{k,1:2}).$$

Because $g(0) \le 0 \le g(1)$ and g is strictly increasing, *bisection* converges to $\tau^\star$ with bracketing on $[0,1]$.

Vectorized bisection (batched) For all items in a batch and all experts/time-bins (indices suppressed):

1. Initialize lo $\leftarrow 0$, hi $\leftarrow 1$.

2. For $s = 1, \dots, S = 20$:
   (a) $\text{mid} \leftarrow (\text{lo} + \text{hi})/2$.
   (b) $v \leftarrow \tilde{F}_{k,\mathbf{x}}(\text{mid}) - t_j$.
   (c) Update $\text{lo} \leftarrow \mathbf{1}_{\{v<0\}} \cdot \text{mid} + \mathbf{1}_{\{v\geq 0\}} \cdot \text{lo}$, $\text{hi} \leftarrow \mathbf{1}_{\{v<0\}} \cdot \text{hi} + \mathbf{1}_{\{v\geq 0\}} \cdot \text{mid}$.
3. Return $\tau^\star \approx (\text{lo} + \text{hi})/2$.

This procedure halves the bracketing interval at each step, so $S = 20$ iterations yield $\approx 10^{-6}$ precision.

Numerical safeguards We clip the denominator $D = F_{k,\mathbf{x}}(1) - F_{k,\mathbf{x}}(0)$ away from 0 when forming $\tilde{F}_{k,\mathbf{x}}$, bound the slopes $a_{k,r}(\mathbf{x}) \in [a_{\min}, a_{\max}]$ (e.g., $a_{\min} = 0.1$, $a_{\max} = 35$), enforce $w_{k,1:2}$ via a softmax, and enforce ordered centers by a stick-breaking parameterization to improve conditioning.

B. Gradients via the Implicit Function Theorem

Let $\tau^\star = \psi_{k,\mathbf{x}}(t_j)$ satisfy $g(\tau^\star; \theta) = 0$ with $g(\tau; \theta) = \tilde{F}_{k,\mathbf{x}}(\tau; \theta) - t_j$. Because $\partial_\tau \tilde{F}_{k,\mathbf{x}}(\tau^\star; \theta) > 0$, the implicit function theorem gives

$$\frac{\partial \tau^\star}{\partial \theta} = -\frac{\partial_\theta \tilde{F}_{k,\mathbf{x}}(\tau^\star; \theta)}{\partial_\tau \tilde{F}_{k,\mathbf{x}}(\tau^\star; \theta)}. \tag{27}$$

Write $F = F_{k,\mathbf{x}}$, $F_0 = F(0)$, $F_1 = F(1)$, and $D = F_1 - F_0$. Using

$$\tilde{F}(\tau) = \frac{F(\tau) - F_0}{D},$$

we obtain

$$\partial_\tau \tilde{F}(\tau) = \frac{\partial_\tau F(\tau)}{D}, \tag{28}$$

$$\partial_\theta \tilde{F}(\tau) = \frac{\partial_\theta F(\tau) - \partial_\theta F_0 - \tilde{F}(\tau)\,(\partial_\theta F_1 - \partial_\theta F_0)}{D}. \tag{29}$$

For $F(\tau) = \sum_{r=1}^{2} w_r\, \sigma(a_r(\tau - c_r))$ we have the elementary partials

$$\partial_\tau F(\tau) = \sum_{r=1}^{2} w_r\, a_r\, \sigma'(a_r(\tau - c_r)), \tag{30}$$

$$\sigma'(z) = \sigma(z)(1 - \sigma(z)), \tag{31}$$

$$\partial_{w_r} F(\tau) = \sigma(a_r(\tau - c_r)), \tag{32}$$

$$\partial_{a_r} F(\tau) = w_r\,(\tau - c_r)\,\sigma'(a_r(\tau - c_r)), \tag{33}$$

$$\partial_{c_r} F(\tau) = -\,w_r\, a_r\, \sigma'(a_r(\tau - c_r)), \tag{34}$$

and the same forms evaluated at $\tau = 0$ and $\tau = 1$ for F_0 and F_1. Substituting (29) into (27) yields closed-form expressions for $\partial \tau^\star / \partial \theta$.

Appendix I. Discovered Patient Groups from a SUPPORT2 Personalized Mixture-of-Experts Model

C_0 **($n = 70$, mid KM)** Younger ARF/MOSF patients (often septic) from the second-lowest income stratum. KM initially tracks C_4 despite the age gap and then diverges later, consistent with later DNR decisions; physiologic severity is not extreme, so outcomes sit in the middle of the cohort.

C_1 **($n = 3$, n/a (tiny))** Very elderly with metastatic lung cancer. The cell size is too small for inference, so we treat any apparent patterns as anecdotal.

C_2 **($n = 173$, mid KM)** Older White patients with a cardiopulmonary/chronic-liver mix (lung cancer/COPD/CHF/cirrhosis). Outcomes are mid-range, worse than C_3 but better than the high-risk groups, consistent with substantial chronic disease burden without the acute derangements seen in C_5/C_7.

C_3 **($n = 47$, best KM)** Middle-aged men with metastatic colon cancer and chronic disease (CHF/COPD/cirrhosis) but ESRD on hemodialysis may explain the creatinine right-tail without acute renal failure; ADLs are low. They show the best vitals (lower HR, good MAP), best pH and P/F ratio, least fever, and physicians also predicted short-term survival. Despite poor glucose control, this group has the best observed survival.

C_4 **($n = 78$, varied KM)** Older Black patients with ARF/MOSF (often septic) and DNR status; the KM curve shows a bimodal/late drop that may reflect abrupt physiologic collapse uninterrupted due to DNR status. Clinicians labeled them "risky in 2 months," and the early KM portion resembles C_0 (with less extreme outcomes) despite the age difference.

C_5 **($n = 221$, worst KM)** Oldest cohort, dementia, coma, frequent malignancy/sepsis, and DNR very early after admission; HR is high, pH is acidotic, creatinine is markedly elevated.

C_6 **($n = 188$, mid KM)** 60-something patients with COPD/CHF/cirrhosis and diabetes from the lowest income stratum. Despite comparatively favorable physiology in places and high predicted survival

in other analyses, outcomes are worse than C_3; elevated HR.

C_7 **($n = 131$, elevated KM)** 50-something patients with ARF/MOSF (± sepsis) who are often coma/intubated from the second-lowest income stratum; creatinine and fever are prominent, and HR is among the higher distributions. Mortality is high and clinicians undercalled 2-month risk relative to C_4, indicating under-recognized early hazard in this phenotype.

Appendix J. Cell-wise z-scores for surfacing cluster-specific categories

For each categorical variable with levels $k \in \{1, \ldots, K\}$ and cluster labels $c \in \{1, \ldots, C\}$, let $n_{c,k}$ denote the count of observations in cluster c having level k. Let row, column, and grand totals be

$$n_c = \sum_k n_{c,k}, \qquad n_k = \sum_c n_{c,k}, \qquad N = \sum_{c,k} n_{c,k}.$$

Under the independence (no-association) model, the expected count in cell (c, k) is

$$e_{c,k} = \mathbb{E}[n_{c,k}] = \frac{n_c n_k}{N}. \tag{35}$$

We compute the adjusted Pearson (Haberman) residuals

$$z_{c,k} = \frac{n_{c,k} - e_{c,k}}{\sqrt{e_{c,k}(1 - \frac{n_c}{N})(1 - \frac{n_k}{N})}}, \tag{36}$$

which standardize observed-expected deviations.

Screening rule We surface "above-background" cluster characteristics by flagging cells with $z_{c,k} > 2$ (and, symmetrically, "below-background" with $z_{c,k} < -2$). In practice we report all cells with $|z_{c,k}| > 2$.

J.1. Flagged attributes by cluster

The following lists the flagged categorical variable levels for each cluster from the SUPPORT2 Personalized MoE model, along with their corresponding z-score values in parentheses. This analysis can also be performed visually using heatmaps of the z-scores as in Figure 6.

Cluster 0 tags:

- death: 0 (value: 5.48)
- hospdead: 0 (value: 4.39)
- dzgroup: ARF/MOSF w/Sepsis (value: 3.05)
- dzclass: ARF/MOSF (value: 2.43)
- income: $25-$50k (value: 5.03)
- ca: no (value: 2.21)
- dnr: no dnr (value: 5.33)
- adlp: 1.0 (value: 2.27)
- sfdm2: no(M2 and SIP pres) (value: 3.22)
- sfdm2: missing (value: 4.15)

Cluster 1 tags:

- dzgroup: Lung Cancer (value: 5.03)
- dzclass: Cancer (value: 3.97)
- ca: metastatic (value: 3.39)
- adlp: 0.0 (value: 3.30)
- sfdm2: no(M2 and SIP pres) (value: 2.42)

Cluster 2 tags:

- hospdead: 0 (value: 6.30)
- dzgroup: Cirrhosis (value: 2.67)
- dzgroup: Lung Cancer (value: 2.35)
- dzgroup: COPD (value: 3.98)
- dzgroup: CHF (value: 4.61)
- dzclass: Cancer (value: 2.60)
- dzclass: COPD/CHF/Cirrhosis (value: 7.49)
- race: white (value: 2.87)
- diabetes: 0 (value: 2.45)
- dnr: no dnr (value: 4.32)
- adlp: 3.0 (value: 2.98)
- adlp: 2.0 (value: 2.12)

Cluster 3 tags:

- sex: male (value: 2.41)
- hospdead: 0 (value: 4.23)
- dzgroup: Colon Cancer (value: 8.94)
- dzgroup: CHF (value: 5.20)
- dzclass: Cancer (value: 5.91)
- dzclass: COPD/CHF/Cirrhosis (value: 3.35)
- ca: metastatic (value: 4.52)
- dnr: no dnr (value: 5.16)
- adlp: 2.0 (value: 4.25)
- sfdm2: no(M2 and SIP pres) (value: 4.13)

Cluster 4 tags:

- death: 1 (value: 5.07)
- hospdead: 1 (value: 6.24)
- dzgroup: ARF/MOSF w/Sepsis (value: 3.13)
- dzclass: ARF/MOSF (value: 3.57)
- race: black (value: 2.09)
- dnr: dnr after sadm (value: 12.72)
- adlp: missing (value: 2.35)
- sfdm2: <2 mo. follow-up (value: 6.99)

Cluster 5 tags:

- death: 1 (value: 8.59)
- hospdead: 1 (value: 18.60)
- dzgroup: ARF/MOSF w/Sepsis (value: 2.16)
- dzgroup: MOSF w/Malig (value: 5.74)
- dzgroup: Coma (value: 7.79)
- dzclass: ARF/MOSF (value: 5.13)
- dzclass: Coma (value: 7.79)
- income: missing (value: 2.00)
- dementia: 1 (value: 2.48)
- ca: yes (value: 4.68)
- dnr: dnr before sadm (value: 3.28)
- dnr: dnr after sadm (value: 15.98)
- adlp: missing (value: 11.33)
- sfdm2: <2 mo. follow-up (value: 16.70)

Cluster 6 tags:

- death: 0 (value: 6.80)
- hospdead: 0 (value: 9.07)
- dzgroup: COPD (value: 3.67)
- dzgroup: CHF (value: 4.39)
- dzclass: COPD/CHF/Cirrhosis (value: 5.19)
- income: $11-$25k (value: 3.10)
- diabetes: 1 (value: 3.55)
- dnr: no dnr (value: 10.58)
- adlp: 1.0 (value: 3.56)
- adlp: 0.0 (value: 12.00)
- sfdm2: no(M2 and SIP pres) (value: 8.86)

Cluster 7 tags:

- death: 0 (value: 2.43)
- hospdead: 0 (value: 4.02)
- dzgroup: ARF/MOSF w/Sepsis (value: 6.86)
- dzclass: ARF/MOSF (value: 7.12)
- income: $25-$50k (value: 2.29)
- dnr: no dnr (value: 5.29)
- dnr: missing (value: 3.45)
- sfdm2: SIP >= 30 (value: 3.12)
- sfdm2: adl>=4 (>=5 if sur) (value: 5.23)
- sfdm2: Coma or Intub (value: 3.45)

Appendix K. Additional Figures

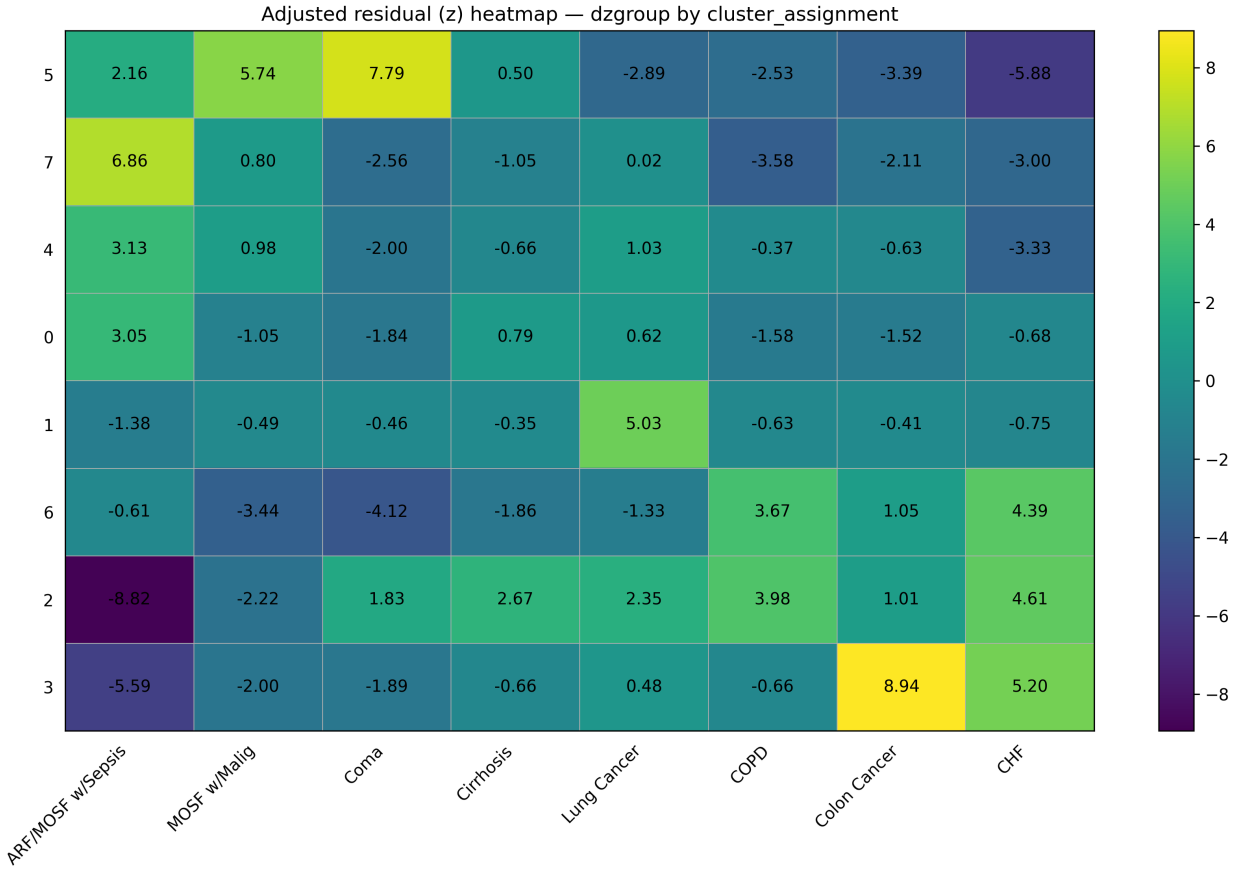


Figure 6: Heatmap of z-scores for the `dzgroup` attribute in SUPPORT2 based on cluster assignments from the Personalized MoE model.

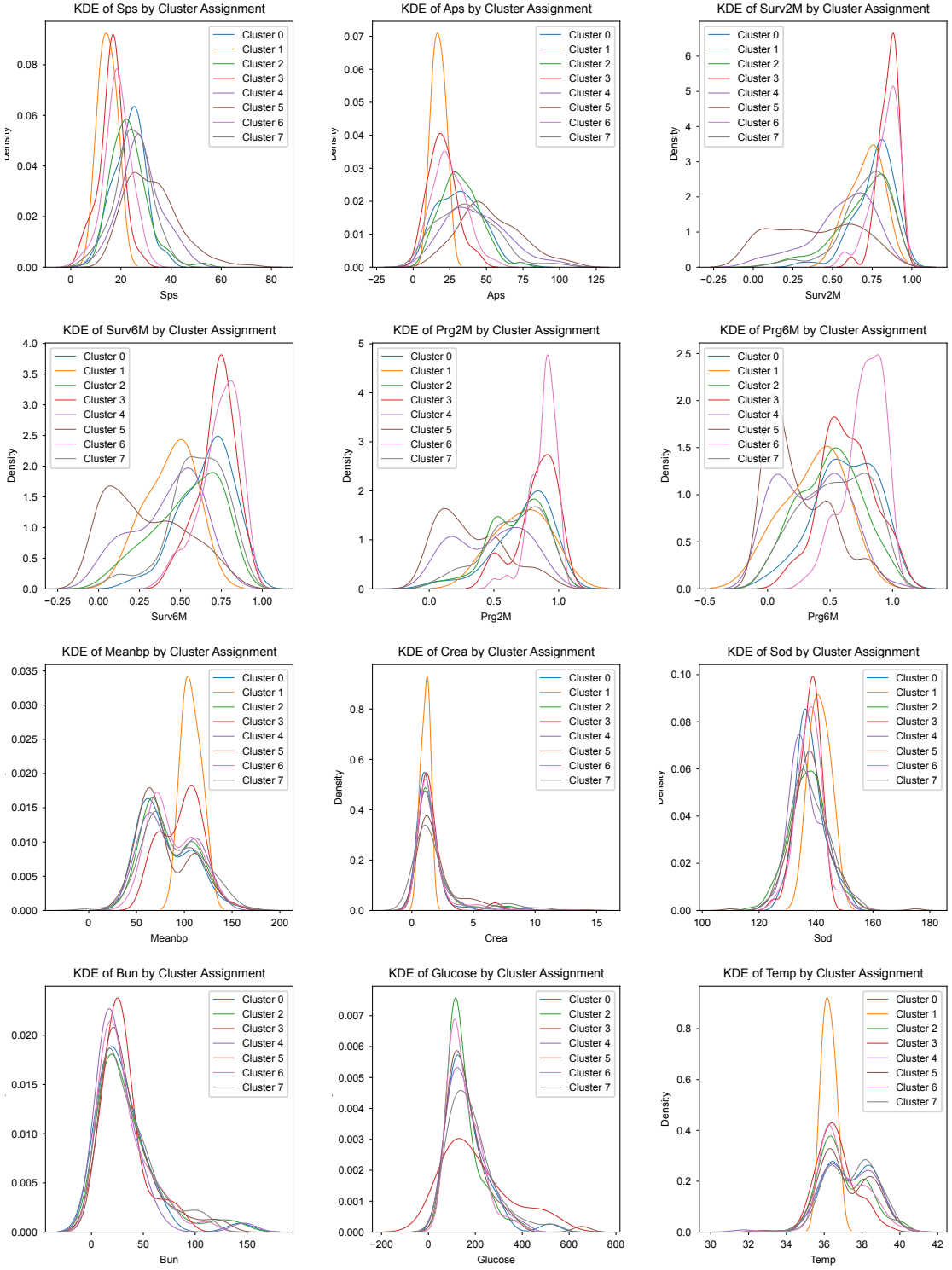


Figure 7: KDE plots for continuous variables by cluster assignment. Refer to data dictionary at https://archive.ics.uci.edu/dataset/880/support2 for variable definitions.

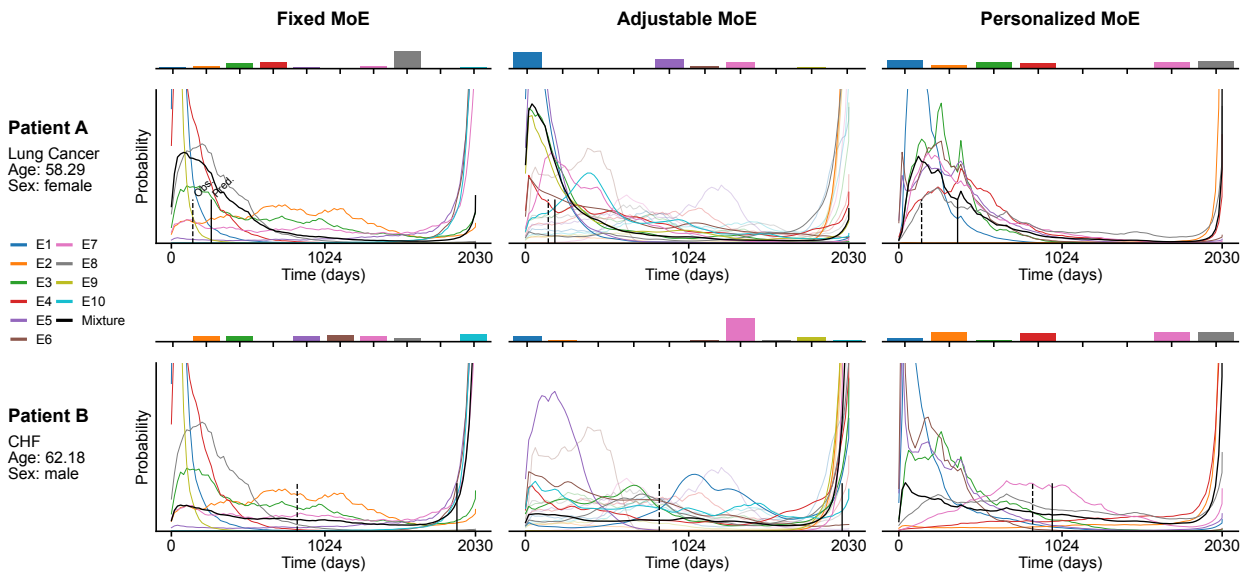


Figure 8: All plots from models trained on SUPPORT2. Left - **Fixed MoE** model event distributions for each expert (faint colors) and adjustments (darker colors), middle - **Adjustable MoE** model event distributions for each expert, and right - **Personalized MoE** model event distributions for each expert.

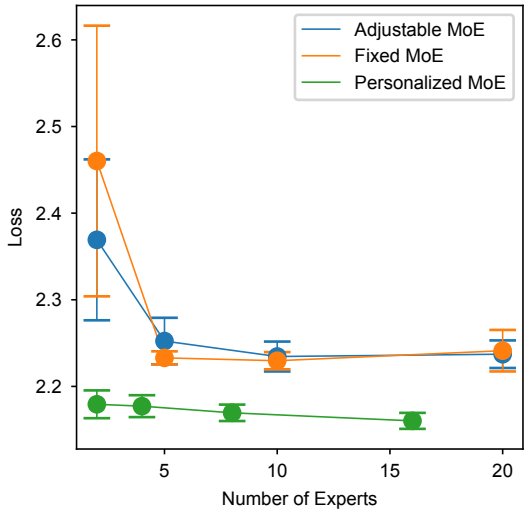


Figure 9: SUPPORT2 expert sensitivity analysis over 5 random seeds varying the number of experts. Test set loss as a function of the number of experts.