Harmonic Mean Point Processes: an Approach to Proportional Rate Error Minimization Across Time for Obtundation Prediction

Yoonjung Kim and Jeremy C. Weiss
Heinz College of Information Systems and Public Policy
Carnegie Mellon University
Pittsburgh, PA 15213
{yoonjungkim,jeremyweiss}@cmu.edu

Abstract

In healthcare, the highest risk individuals for morbidity and mortality are rarely those with the greatest modifiable risk. By contrast, many machine learning formulations implicitly attend to the highest risk individuals. We focus on this problem in point processes, a popular modeling technique for the analysis of the temporal event sequences in electronic health records (EHR) data with applications in risk stratification and risk score systems. We show that optimization of the log-likelihood function also gives disproportionate attention to high risk individuals and leads to poor prediction results for low risk individuals compared to ones at high risk. We characterize the problem and propose an adjusted log-likelihood formulation as a new objective for point processes. We show that the proposed formulation is a weighted sum of segmented likelihood contributions that allows proportionate attention given data generated from a stochastic model corresponding to a harmonic mean estimator. We demonstrate the benefits of our method in simulations and in EHR data of patients admitted to the critical care unit for intracerebral hemorrhage.

1 Introduction

Clinical forecasting is a central task for prognostication in populations at risk for downstream morbidity and mortality. When followup is incomplete and right-censorship of data occurs, survival analysis models are often preferable to binary classification for long-term prognostication due to their ability to mitigate selection bias associated with censorship. Many models exists for the survival analysis task, including Cox, Aalen, accelerated failure time models, random survival forests, and so on. Point processes are a natural generalization of the survival analysis task when modeling data in the presence of repeated outcome events, where rate modeling is not limited to conditional rate estimation, where the condition is event-free survival.

In rate models, characterization of proportional errors in rate is often an objective of primary interest, but its minimization is not straightforward because ground truth rates are unobserved. Indeed, the long-standing success of the Cox proportional hazards model is a demonstration of the interpretive value of proportional rate estimation. Despite this, the objective function specified for many survival models do not seek to minimize proportional error in rate, including that of Cox.

We approach this problem through the formulation of the optimization objective function. Our analysis illustrates that the standard likelihood function attends to individuals at highest risk, potentially at the cost of modeling proportional rates in low-risk individuals poorly. This characterization provides us a way to attempt to mitigate the mis-attention and will result in a reweighting scheme to fairly attend

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

to all individuals with respect to proportional rate misspecification. Practically, this will result in a fairness-variance trade-off, as the models will suffer from having high variance from low effective sample sizes.

One natural solution to the above problem is to optimize for the proportional predictive error. Accelerated failure time (AFT) models accomplish this by log transforming time so that minimizing mean squared error in log space corresponds to minimizing multiplicative errors in the original space. This approach works for single event analyses, but fails to extend to repeated event models. In particular, AFT models require a specification of time $t = 0$ so that the log transformation is well-defined. Time $t = 0$ specification is problematic in time-varying and nowcasting analysis, as noted in (Miscouridou et al., 2018) and (Weiss, 2018), because training time $t = 0$ must be specified while the modeler may want to vary test time $t = 0$ or model repeated events. Attempts to duplicate training samples longitudinally by varying train time $t = 0$ lead to samples that violate the parametric and or semi-parametric model assumptions and result in poor parameter estimates and poor predictive performance. Because of this problem, it is unclear how to naturally extend such models to recurrent event and multitask settings. In these settings, larger ranges of rates are often modeled, and this magnifies the problem of misplaced attention.

In this work, we propose a model to minimize proportional errors in rates in settings of anytime prediction and recurrent event modeling. Our model reweights the log likelihood according to the inverse of the predicted rate, so that the likelihood attends to high rates and low rates proportionally fairly. We show that as a result of our objective specification that we recover a harmonic mean risk estimator. Because rates are expectations of events per unit time, *i.e.* arithmetic means, the harmonic mean estimator underestimates the risk in face of random effects or frailty. Nonetheless, because we do not have oracle access to the true rates, a point prediction method does not result in such underestimation. Instead, empirically we must trade off between reweighting fairly with maintaining adequate effective sample size for stable risk prediction. Finally we demonstrate in simulations and in prediction of neurological deterioration among patients admitted for intracerebral hemorrhage (ICH) that our method empirically produces informative risk assessments in low rate regimes.

1.1 Related Work

The literature on the use of machine learning for survival analysis or point processes is large, and we limit our discussion to closely-related and recent works in the space. While our method is general for methods that optimize for the survival objective function, we devise two models that extend previous approaches: that of (1) piano roll embeddings, an LSTM-variant (Dong et al., 2018), and (2) wavelet reconstruction networks (Weiss, 2018). The former is meant to forecast near term future music chords over continuous time, which can be modified to make survival predictions according to the maximum likelihood or harmonic mean objective functions. The latter jointly trains relative-time wavelet kernel functions and the function that combines them to represent a survival function on absolute time, but uses the maximum likelihood survival objective. Similarly, the approach could be used in training other recent survival and point process frameworks, including Jing and Smola (2017); Lee et al. (2018); Weiss (2019). Several recent works have adopted alternative objective functions, including (Avati et al., 2018) and (Chapfuwa et al., 2018). Another approach is to fuse together binary classification predictions across time. Our method may help illustrate the effectiveness of this approach, in that this multi-task prediction formulation attends implicitly to low-risk examples by ignoring those examples where events have already occurred in the large time-to-event classifiers (Yu et al., 2011). However, unlike these methods, our method provides attention to low-risk persons and regions in the repeated event setting. Our finding that the oracle estimator is a harmonic mean estimator suggests that, if one were to extend our model to output distributions of risk, *i.e.*, frailty models, for example by using a conditional variational autoencoder or generative adversarial network, that the reweighting scheme would underestimate average risk. A rich literature on frailty models exists, though the methods described typically use small parametric models. Finally, several other machine learning models have leveraged the time rescaling theorem (Pillow, 2009; Gerhard and Gerstner, 2010), but none formulating or motivated by a low-risk rate estimation.

2 Background

Let Y be the event we want to model over time across N samples. Let the event times be the sequence t_{in} for $i \in \{1, \ldots, T_n\}$ for $n = \{1, \ldots, N\}$ over a period of interest $[0, \tau_n]$. We are interested in modeling the rate function:

$$\lambda(t|\cdot) = \lim_{h \to 0} \frac{P(t < T < t + h | T > t, \cdot)}{h} = \frac{f(t|\cdot)}{S(t|\cdot)},$$

where $\{\cdot\}$ varies by model and represents the information or data to use in modeling λ. The probability density function and survival function are given by f and S. Given parameters Θ, the survival log likelihood is given by:

$$\text{LL}(\mathbf{X}|\Theta) = \sum_{n=1}^{N} \Big(\sum_{i=1}^{T_n} \log \lambda_n(t_{in}|\cdot) - \int_0^{\tau_n} \lambda(t|\cdot) dt \Big) \tag{1}$$

For Cox processes, the form is $\lambda(t; x) = \lambda_0(t) e^{w^T x}$ where $\lambda_0(t)$ is a nuisance function, w is the parameter vector, and x is a feature vector per example. Aalen additive models have form $\lambda(t; x) = \lambda_0(t) + w(t)^T x$ for functions $\lambda_0(t)$ and $w_i(t)$ for $i \in \{1, \ldots, k\}$ for feature vector x of length k. Hawkes processes condition on the history $\mathcal{H}(t)$ up until time t: $\lambda(t|\mathcal{H}(t)) = \lambda_0(t) + \sum_{i=1}^{T_n} g(t - t_i) \mathbb{1}(t_i < t)$, where $g(\cdot)$ is a kernel function (usually an exponential decaying function) and $\mathbb{1}(\cdot)$ is the indicator function.

Next we establish the relationship between rescaled time where a single Poisson process with rate 1 and our original time, where λ is defined. This comes from the time rescaling theorem.

Time rescaling theorem. Given the rate function λ, define Λ the cumulative hazard function: $\Lambda(t|\cdot) = \int_0^{\tau_n} \lambda(t|\cdot) dt$. For the realization of a sequence of events from $\lambda(t|\cdot)$ with times $\{u_1, \ldots, u_k\}$ and $\Lambda(\tau|\cdot) < \infty$, the sequence $\{\Lambda(u_1|\cdot), \ldots, \Lambda(u_k|\cdot)\}$ is distributed according to a unit rate Poisson process (Meyer, 1971; Ogata, 1981).

Details of the proof can be found in (Brown et al., 2002). The implication of the theorem is that if we could model the conditional intensity correctly, the intervals between rescaled times follow exponential distribution with rate 1.

As an illustration, maximizing the equation 1 and log-likelihood in the rescaled time having Poisson process with rate 1 yield the same conditional intensity assuming a piecewise constant functional form where the intensity values are locally constant. Any intensity function could be approximated as piecewise constant by assigning constant intensity values to infinitesimal time intervals. The simplest case that demonstrates the equivalence of the two maximization problems would be a constant intensity λ with one sample. Let's assume n events occurred during time τ for the sample. We get the estimator of λ, $\hat{\lambda}$ as follows:

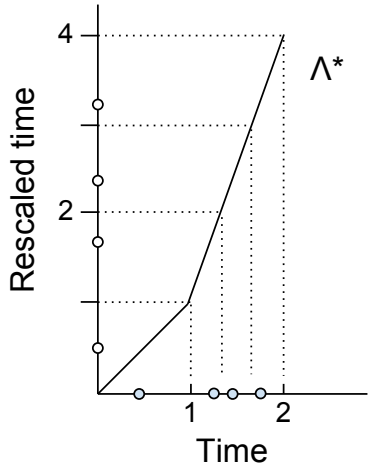
$$\hat{\lambda}_1 = \operatorname*{argmax}_{\lambda} LL(X|\lambda) = \operatorname*{argmax}_{\lambda} \sum_{i}^{n} \log \lambda - \int_0^{\tau} \lambda dt = \operatorname*{argmax}_{\lambda} n \log \lambda - \lambda \tau = \frac{n}{\tau} \tag{2}$$

$$\hat{\lambda}_2 = \operatorname*{argmax}_{\lambda} \log \left(\frac{e^{-\lambda\tau} (\lambda\tau)^n}{n!} \right) = \operatorname*{argmax}_{\lambda} -\lambda\tau + n \log \lambda - \log(n!) = \frac{n}{\tau} \tag{3}$$

Equation 2 computes the the estimator $\hat{\lambda}_1$ with the log-likelihood, while equation 3 computes the the estimator $\hat{\lambda}_2$ with the log-likelihood in the rescaled time. In the bottom line, we get the same estimator $\hat{\lambda} = \hat{\lambda}_1 = \hat{\lambda}_2$ with both equations. This can be extended to smooth functions under regularity conditions by considering the upper limit as the step size goes to zero.

3 Method

From the time rescaling theorem, it is straightforward to observe that the relative contributions to the likelihood of each time interval is proportional to the rate within that interval, *i.e.*, if we care about each individual's risk in a time unit equally, then we could consider decreasing the likelihood contributions in proportion to the rate. In other words, our procedure will seek to nullify, partially



Algorithm 1: Harmonic Mean Point Processes

Result: ALL-trained model
Temporal network $F : X \mapsto [0, \infty)$;
Attention coefficient γ, stability factor ϵ ;
while *training* **do**
  $\hat{\lambda}(t'_{j,k-1}) = F(X_j)$ piecewise on $[t'_{j,k-1}, t'_{j,k}) \; \forall k \in K$;
  Copy then detach $\hat{\lambda}(t'_{j,k-1}) \; \forall j, k$;
  $\text{ALL}_{j,k} = \text{LL}_{j,k} / \big(\hat{\lambda}(t'_{j,k-1})^{\log_{10} \gamma} + \epsilon\big)$;
  ALL.sum().backward();
  optimizer.step();
end

Figure 1: Rescaled time illustrating a unit of $\lambda = 1$ has three times less likelihood weight than one unit of $\lambda = 3$, evidenced by its proportion of the length of the y-axis (left). Pseudocode for HMPPs (right).

or fully, the proportional factor of likelihood attention given to higher rates. We call this approach optimization of the adjusted log likelihood, which is illustrated in Figure 1.

$$ALL(\mathbf{X}|\Theta) = \sum_{n=1}^{N} \Big(\sum_{i=1}^{T_n} \frac{\log \lambda(t_{in}|\Theta)}{\lambda^*(t_{in})} - \int_0^{\tau_n} \frac{\lambda(t|\Theta)}{\lambda^*(t)} dt \Big) \tag{4}$$

where $\lambda^*(t)$ is the ground truth intensity at time t. By assuming $\lambda(t)$ and $\lambda^*(t)$ are piecewise constant, we can view the adjusted log-likelihood as the weighted sum of log-likelihood contributions. Suppose we divide the time interval $(0, \tau_j]$ into K sub-intervals where K is a significantly large number so that $\lambda^*(t)$ is constant within any sub-interval. That is, with $0 = {t'}_{j,0} < {t'}_{j,1} < \cdots < {t'}_{j,K} = \tau_j$, $\lambda^*(t)$ is constant for $t \in ({t'}_{j,k-1}, {t'}_{j,k}]$ and all $k = 1, \ldots, K$. Then,

$$ALL(\mathbf{X}|\Theta) = \sum_{j=1}^{m} \left[\int_0^{\tau_j} \frac{\log \lambda(t|\Theta)}{\lambda^*(t)} dN(t) - \int_0^{\tau_j} \frac{\lambda(t|\Theta)}{\lambda^*(t)} dt \right] \tag{5}$$

$$= \sum_{j=1}^{m} \sum_{k=1}^{K} \left[\int_{{t'}_{j,k-1}}^{{t'}_{j,k}} \frac{\log \lambda(t|\Theta)}{\lambda^*(t)} dN(t) - \int_{{t'}_{j,k-1}}^{{t'}_{j,k}} \frac{\lambda(t|\Theta)}{\lambda^*(t)} dt \right] \tag{6}$$

$$= \sum_{j=1}^{m} \sum_{k=1}^{K} \frac{1}{\lambda^*({t'}_{j,k})} \left[\int_{{t'}_{j,k-1}}^{{t'}_{j,k}} \log \lambda(t|\Theta) dN(t) - \int_{{t'}_{j,k-1}}^{{t'}_{j,k}} \lambda(t|\Theta) dt \right] \tag{7}$$

is the sum of log-likelihood contributions in time intervals $({t'}_{j,k-1}, {t'}_{j,k}]$ weighted by the reciprocal of the ground truth intensity at ${t'}_{j,k}$ for every j and k. The similarity is apparent when it is compared to the similar form of standard log-likelihood:

$$LL(\mathbf{X}|\Theta) = \sum_{j=1}^{m} \left[\int_0^{\tau_j} \log \lambda(t|\Theta) dN(t) - \int_0^{\tau_j} \lambda(t|\Theta) dt \right] \tag{8}$$

$$= \sum_{j=1}^{m} \sum_{k=1}^{K} \left[\int_{{t'}_{j,k-1}}^{{t'}_{j,k}} \log \lambda(t|\Theta) dN(t) - \int_{{t'}_{j,k-1}}^{{t'}_{j,k}} \lambda(t|\Theta) dt \right] \tag{9}$$

Therefore, we can weight each interval's log likelihood by the inverse of the oracle rate to get the adjusted log likelihood.

3.1 Oracle approximation

Without access to λ^*, however, we must resort to approximation of the reweighting. One choice for λ^* is our current estimate $\hat{\lambda}$. However, this could lead to unstable weightings because a single example could dominate the weight distribution. To address this fairness-variance tradeoff, we introduce the attention coefficient γ and stability factor ϵ to help stabilize the weights. Pseudocode in Figure 1 (right) illustrates the training procedure and the stabilization modification. We call our method harmonic mean point processes (HMPP) because if the oracle is known and doubly stochastic (frail), then the estimates we get from the training procedure are harmonic mean estimates of the rate distribution. Note that in practice when we use an approximation, the denominator must be copied and detached from the computation graph so that the graph of which λ is a part is not further connected by the current model's predictions $\hat{\lambda}$.

Figure 2 demonstrates the reweighting achieved with different choices of attention coefficient γ and stability factor ϵ. To achieve equal rescaled-time weighting, γ must be set to 10, corresponding to 10-fold increased weight per 10-fold decrease in risk: the blue horizontal line. However, the number of effective samples may become very small, shown by the number of effective samples (per 1) for several common distributions. To avoid this, γ and or ϵ can be chosen to flatten the reweighting distribution. In practice the predicted distribution is implicit and potentially unstable, and using domain knowledge to set ϵ near to the lowest rate expected to be found will mitigate the instability while still attending to the low risk individuals.

4 Experimental Setup

We test our method in two simulations, where we have access to ground truth rates, and in application to a health setting, where we illustrate important factors and effects of our approach. In all cases, we are comparing our objective versus the standard variant, and call the models Harmonic Mean Point Processes (HMPP, ours) and Maximum Likelihood Point Processes (MLPP, comparison). We use this labeling across multiple models and domains which we describe next.

4.1 Simulations

To test our ability to accurately determine the rates of low risk individuals, we developed a singly- and a doubly- stochastic univariate model with rates varying by 4-6 orders of magnitude (10^4 to 10^6 fold variation in rates). In each case, events are sampled for 10 units of time according to a sample-specific fixed rate λ^*, where λ^* is drawn from a truncated, base 10 exponential between 10^{-2} to 10^2. For the singly-stochastic model, we sample events according to an exponential with rate λ^*, and for the doubly-stochastic model, we sample events according to an exponential with rate $\lambda^* 10^u$ for $u \sim \text{Uniform}(-2, 0)$. Then, a time-stamped sequence is produced, containing tuples of (id, time, event, value) features, with the first tuple containing (id, 0, rate, λ^*). Thus the recurrent networks can learn from the value, or it can draw from the timing of previous events in the sequence.

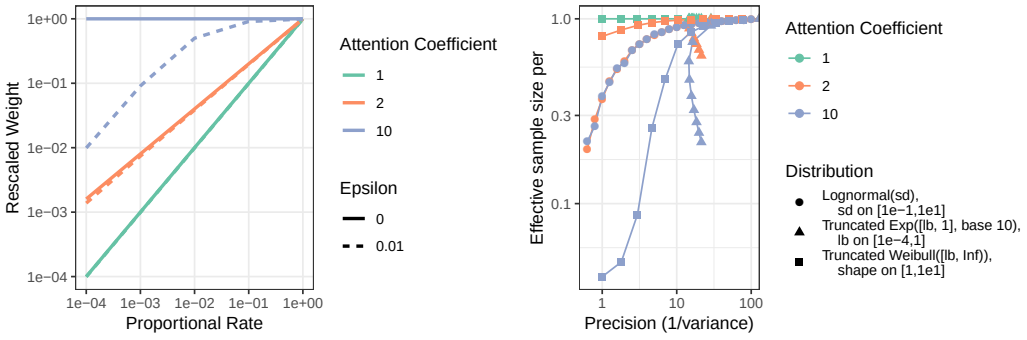


Figure 2: Rescaled-space weight based on relative predicted weight, attention coefficient γ (MLPP : $\gamma = 1$), and attention stability factor ϵ (left). Effective sample size per example under distributional assumption of rates (right).

While these simulations are simple, it will demonstrate the point that recurrent networks without the use of the adjusted log likelihood produce substantial overestimates for a large fraction of low-risk samples. We sample 10,000 training and test examples for each simulation. We use an embedding LSTM architecture for the simulations (Dong et al., 2018) and provide details in the Appendix.

4.2 Application: obtundation in intracerebral hemorrhage

We also apply our method to real data of neurological decline during critical care admissions for intracerebral hemorrhage (ICH). We provide a brief description to motivate the problem and describe the experimental setup for this application.

Intracerebral hemorrhage is a life-threatening extravasation of blood outside the vessel wall due to a tear or rupture that results in a hematoma or accumulation of blood, which then presses upon the soft tissue of the brain causing neuronal damage. Mortality rates are 40% at 1 month post diagnosis. In these individuals, frequent monitoring of neurological status is essential. The Glasgow Coma Score (GCS) is a score based on physical exam that provides an indicator for progression and recovery. For ICH, GCS is a primary indicator in several ways. First, mortality stratification is conducted, *e.g.* a hematoma larger than 60cm^2 and a GCS score below 8/15 has a 1-month mortality rate of 90%. Similarly, decreasing GCS is a primary risk factor for poor long-term prognosis. Second, some protocols use GCS below 8 as a threshold for intracranial pressure monitoring and intubation. As a result, the rate of GCS testing in this population is high on average.

For many individuals, however decreasing GCS is unlikely to occur. Reasons for this could include: a trajectory of neurological recovery, strong sedation has been given, or the patient is being held for additional non-neurological reasons. In these individuals, while a decreasing GCS may be less frequent, a decreasing GCS may be more significant. A decreasing GCS is also relatively underweighted by learning algorithms that optimize for "easier" GCS predictions such as after sedative administration. Modeling clinical risks from electronic health records (EHRs) may support clinicians predicting decreasing GCS scores. We used data from MIMIC III v1.4 (Johnson et al., 2016), an EHR housing critical care data on 40,000 individuals. Of those, 1,010 had a primary diagnosis of ICH and were considered as members of our cohort. Chart, laboratory, medications, vitals, procedures, and demographics tables were extracted as time-varying features for nowcasting GCS decreases. GCS scores were recorded in the chart table in two versions depending on the vendor, one by the component scores Eyes, Verbal, and Motor, and the other as an aggregate score. We defined a decrease in GCS to be a decrease of any score not a result of intubation, or a first GCS below 8 (an obtunded state indicative of poor outcomes). GCS readings inside the critical care unit were considered only. Per individual, events within the first ICH encounter only were used. Figure 3 depicts the study characteristics and approach.

We use wavelet reconstruction networks (WRNs) with the objective function modification to accommodate the adjusted log likelihood. The details of the architecture are provided elsewhere (Weiss, 2018), but we provide a brief description. WRNs take as inputs the same 4-tuples as above but instead of embedding events and values in bins for a recurrent architecture, they follow a Hawkes process approach where each event induces a function (a wavelet reconstruction) over time, and these functions are reduced from many to produce a single non-negative rate function. WRNs have performed well in health settings and hold the advantage that, unlike recurrent and dilated convolutional architectures, the time bins need not be pre-specified.

We investigate the performance of the algorithms using inspection of calibration and variable importance plots. Notably, two common evaluation measures are not appropriate here, including: (1) log likelihood, which attends to high risk individuals and is therefore not central for low-risk prediction and (2) concordance (c-) statistic, which is not well defined for recurrent event processes. Instead, we argue that calibration plots both enable visualization of discrimination through risk stratification (order and magnitude) as well as calibration through predicted-versus-expected rate comparisons. We construct calibration plots using the ordering given by the algorithm predictions, which illustrates discriminative ability in the algorithms ability to stratify groups across the spectrum of risk (an alternative is to order by ground truth rates where the specific expressed goal of assessment is calibration). In applications to real data, we cannot access ground truth, so we order by and use equal quantiles with respect to the predictions.

Feature	n=1,010
Age	72 [59, 81]
Gender	
Female	466 (0.46)
Male	544 (0.54)
Ventricular shunt	152 (0.15)
GCS	10 [7, 14]
Decreased GCS, count	3,119
Decreased GCS, rate	1 event / 1.5 days

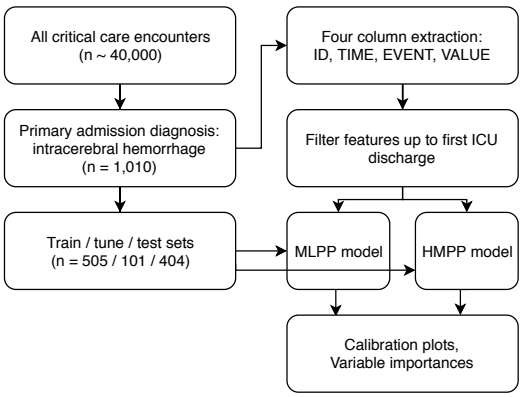


Figure 3: Characteristics of the ICH study population (median, [IQR] or count (%)) and flowchart of the ICH analysis.

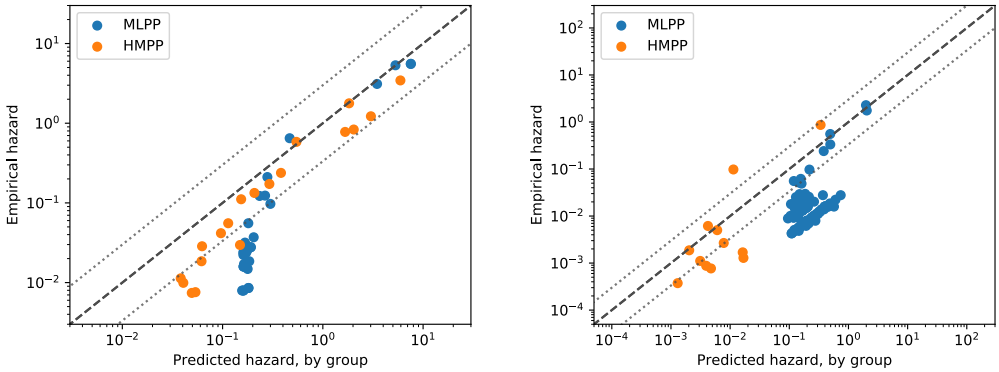


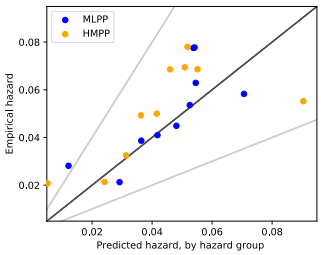
Figure 4: Calibration plots comparing HMPP and MLPP in simulation, singly stochastic (left) and doubly stochastic (right) processes. HMPP predicts groups of individuals with rates an order of magnitude smaller. For fixed effects this comes from improved calibration; for random effects, this comes from discrimination between the low-low and low risk individuals.

Finally we use variable importance plots to illustrate which variables are central to the algorithm's predictions. Since permutation-based variable importance is not available for recurrent event models, we use regularization-based variable importance, which ranks features in importance based on the size of the loss suffered due to regularization on that variable. This typically requires a standardization step for each variable, however, WRNs use wavelet mappings on the value and time distributions that effectively standardize each variable automatically. Variable importance is of interest for our use case in that it can highlight the interesting finding that variables identified as important in high-risk prediction may be a very different set than those important in low-risk prediction.

We conduct training for 50 epochs using the Adam optimizer with a learning rate of 10^{-3} and a batch size of 8. The reweighting at each training step also makes across-step ALL comparisons not meaningful. Therefore, when choosing early stopping points, we use the tune set log likelihood. We also used the tune set performance in our search over the following hyperparameters: $\gamma \in \{10, 2\}$ (ICH only), $\epsilon \in \{0, 10^{-2}\}$ (ICH only), L1 regularization (LASSO) coefficient in $\{10^{-2}, 10^{-3}, 10^{-4}\}$, and L2 regularization (ridge) coefficient in $\{0, 10^{-2}\}$. Our implementation is in PyTorch v1.0.

5 Results

Figure 4 demonstrates the benefit of our method in simulations. In the singly-stochastic model (left figure), both the HMPP and MLPP approaches discriminate risk across the spectrum, illustrated by the (approximately) monotonic curves. However, the MLPP method never predicts hazards lower



HMPP	MLPP	MLPP (HMPP par.)
IV solution	Nasogastric fluid	Foley catheter
Orogastric fluid	Functional fibrinogen	Osmolality:blood
Osmolality: blood	Urine white blood cell count	Temperature
Vancomycin level: blood	Osmolality: blood	Functional fibrinogen
IV dextrose in water: D5W	IV drip labetalol	IV phenytoin
Total cholesterol	Vancomycin level: blood	Phenytoin: blood
Lymphocyte count: csf	Urine red blood cells	IV normal saline
Urine sodium	Basophil count:blood	IV sterile water added
IV normal saline	Urine ketones	PO metoprolol

Figure 5: (Left) Calibration plots for HMPP and MLPP for GCS score decrease prediction. Lower rate groups are identified in HMPP, at the cost of higher variability in predictions. (Right) Top 10 variable importances for HMPP (left), MLPP (right), and MLPP with HMPP regularization parameter settings (bottom). Note the minimal overlap in the lists, and also note the smaller losses in HMPP, illustrative of earlier stopping during HMPP training.

than 0.2 for any group, despite many of those groups having empirical rates near 0.01. By contrast, the HMPP method straightens the low-risk tail and makes more accurate prediction in low risk individuals. At the same time, risk predictions for high risk individuals are similar in quality. In this case, the range of predicted risks from HMPP was half an order of magnitude larger than MLPP.

For the doubly-stochastic model where the formulation includes frailty, the performance of HMPP and MLPP diverges further. In particular, Figure 4 right shows that in the face of random effects that vary the rate ranges by 100-fold, MLPP focuses on the high end of the random effect distribution and HMPP the low end. HMPP identifies groups of individuals with empirical rates an order of magnitude smaller. It also identifies groups at larger rates, but appears to underestimate the rates for these individuals. This could be due to overfitting of the training data leading to erroneously low predictions on the test set. Nonetheless, HMPP detects low risk individuals in this setting whereas MLPP does not acknowledge their low rates, instead limiting all predictions to greater than 0.1.

In the ICH study, the hyperparameters chosen were an elastic net formulation L1 and L2: 10^{-2} with $\gamma = 10$ and $\epsilon = 0.01$, suggesting the model is constrained by limited sample size. In this case, the HMPP and MLPP models produce similar predictions as shown in Figure 5, with MLPP making more conservative predictions and with small empirical risk differences, and HMPP making more variable predictions with larger empirical risk differences. Discriminatively, the C statistic among the lowest quartile was 0.68 (0.62-0.74, bootstrap CI 95%) and 0.66 (0.60, 0.72) for HMPP and MLPP respectively. Thus, while HMPP does identify lower risk groups, the small sample size limits the interpretation.

Even in this small data setting where the low-risk group is not newly identified by the method, we can look at the variable importance plots to demonstrate a marked difference in result. Figure 5 (right) shows HMPP and MLPP variable importances for the top 10 features. Two features overlap, osmolality and vancomycin levels. For the rest, the HMPP model is concerned with lab tests and intravenous solution choices and quantities while the MLPP model is concerned about urine and clotting lab tests. It could be that the need for stability in the form of increased regularization and early stopping could be necessary for the HMPP model, so we additionally computed the variable importances for the MLPP model with the same regularization settings, the top ten variables are also shown, which share no increased overlap with the fair model. This illustrates that the factors that influence proportional risk across the risk spectrum may differ substantially from those obtained from simple likelihood optimization which attends to high risk.

6 Conclusion

Our work demonstrates a new tool to make risk predictions in low-risk populations, when the population being studied possesses members with risk varying across orders of magnitude. We provide a formulation that exhibits how to attend equally across risk, and provide an algorithm and guidance to trade off fair attention with variance from reweighting. Importantly, our method detects individuals an order of magnitude lower than predictions made by optimization with the log likelihood and deep network–the combination of two popular approaches. We further illustrate

implications of attending to low-risk individuals in the variable importances reported under each optimization. This difference may have important applications in suggesting risk factors that are stratum-specific, which can provide guidance in personalized decision making. Future work will include explicit characterization of the proportionate attention-variance tradeoff which could provide alternative approximations to the oracle rate with desirable properties.

References

Avati, A., Duan, T., Jung, K., Shah, N. H., and Ng, A. (2018). Countdown regression: Sharp and calibrated survival predictions. *arXiv preprint arXiv:1806.08324*.

Brown, E. N., Barbieri, R., Ventura, V., Kass, R. E., and Frank, L. M. (2002). The time-rescaling theorem and its application to neural spike train data analysis. *Neural computation*, 14(2):325–346.

Chapfuwa, P., Tao, C., Li, C., Page, C., Goldstein, B., Duke, L. C., and Henao, R. (2018). Adversarial time-to-event modeling. In *International Conference on Machine Learning*, pages 734–743.

Dong, H.-W., Hsiao, W.-Y., Yang, L.-C., and Yang, Y.-H. (2018). Musegan: Multi-track sequential generative adversarial networks for symbolic music generation and accompaniment. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Gerhard, F. and Gerstner, W. (2010). Rescaling, thinning or complementing? on goodness-of-fit procedures for point process models and generalized linear models. In *Advances in neural information processing systems*, pages 703–711.

Jing, H. and Smola, A. J. (2017). Neural survival recommender. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, pages 515–524. ACM.

Johnson, A. E., Pollard, T. J., Shen, L., Li-wei, H. L., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., and Mark, R. G. (2016). Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:160035.

Lee, C., Zame, W. R., Yoon, J., and van der Schaar, M. (2018). Deephit: A deep learning approach to survival analysis with competing risks. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Meyer, P.-A. (1971). Demonstration simplifiee d'un theoreme de knight. In *Séminaire de probabilités v université de strasbourg*, pages 191–195. Springer.

Miscouridou, X., Perotte, A., Elhadad, N., and Ranganath, R. (2018). Deep survival analysis: Nonparametrics and missingness. In *Machine Learning for Healthcare Conference*, pages 244–256.

Ogata, Y. (1981). On lewis' simulation method for point processes. *IEEE Transactions on Information Theory*, 27(1):23–31.

Pillow, J. W. (2009). Time-rescaling methods for the estimation and assessment of non-poisson neural encoding models. In *Advances in neural information processing systems*, pages 1473–1481.

Weiss, J. C. (2018). Clinical risk: wavelet reconstruction networks for marked point processes.

Weiss, J. C. (2019). On microvascular complications of diabetes risk: development of a machine learning and electronic health records risk score.

Yu, C.-N., Greiner, R., Lin, H.-C., and Baracos, V. (2011). Learning patient-specific cancer survival distributions as a sequence of dependent regressors. In *Advances in Neural Information Processing Systems*, pages 1845–1853.

A Architecture

The architecture is given in Figure A.1. The idea is to use a LSTM (piano roll) embedding architecture, where any number of events with or without values can first be embedded as line and point embeddings respectively, and then the embedded signals are captured in a group embedding and passed into LSTM time steps. This architecture facilitates flexible parsing of long format data typical of marked point processes, such as that of digital orchestral music, or that of medical event streams. Categorical events are treated as multiple point events, and point events are embedded as points. Real-valued events are embedded based on their value, and so the event's value domain corresponds to a line embedded as a 1-dimensional manifold. These embedded vectors are then further embedded as a group based on their timestamps, and fed into an LSTM that outputs non-negative rate predictions. We use times steps of unit length with 10 steps in total.

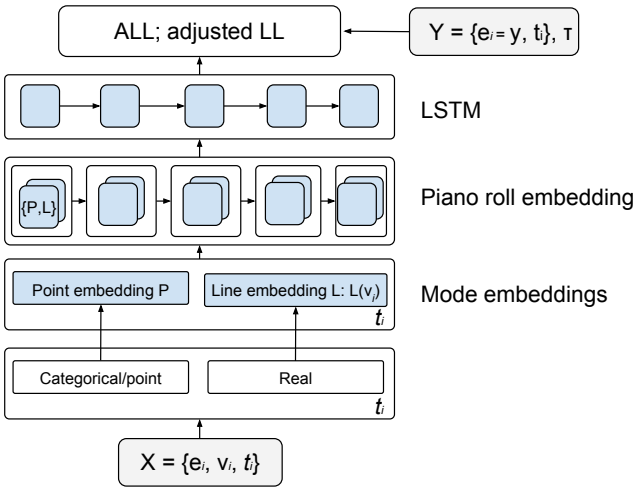


Figure A.1: LSTM embedding architecture used for simulations.