Machine learning and clinical risk: wavelet reconstruction networks for marked point processes

Jeremy C. Weiss
Heinz College of Information Systems and Public Policy
Carnegie Mellon University
Pittsburgh, PA, United States
jeremyweiss@cmu.edu

Abstract

Timestamped sequences of events, pervasive in domains with data logs, *e.g*, health records, are often modeled as point processes with rate functions over time. Leading classical methods for risk scores such as Cox and Hawkes processes use such data but make strong assumptions about the shape and form of multivariate influences, resulting in time-to-event distributions irreflective of many real world processes. Recent methods in point processes and recurrent neural networks capably model rate functions but may be complex and difficult to interrogate. Our work develops a high-performing, interrogable model. We introduce wavelet reconstruction networks, a multivariate point process with a sparse wavelet reconstruction kernel to model rate functions from marked, timestamped data. We show they achieve improved performance and interrogability over baselines in forecasting adverse events and scheduled care visits in patients with diabetes.

1. Introduction

Clinical risk scores are commonly used analytic devices in health care. There are risk scores for predicting strep throat from sore throats (Centor et al., 1981), mortality from vital signs (Gardner-Thorpe et al., 2006), heart attacks from routine clinic visits (D'Agostino et al., 2008), and many more. Policy is implemented around these risk scores, from rates of reimbursement to physician compensation (Asch et al., 2015), and early warning alerts based on risk scores are associated with reduced mortality (Seymour et al., 2017).

A transition is occurring in risk score data collection, from the classical framework of running clinical studies with panel or time series collection protocols, to a newer framework involving digital storage of timestamped events in log files. Electronic health records are the new framework in medicine and are used in conjunction with machine learning algorithms to create personalized risk scores. Figure 1 depicts the two common pathways for developing risk scores side by side, with classic and atemporal models involving extraction, imputation, and learning following the right path (Cox, 1972; Van Buuren & Oudshoorn, 1999), and popular temporal machine learning models following the left (Hawkes, 1971; Lipton et al., 2015; Xu & Zha, 2017; Futoma et al., 2017; Rajkomar et al., 2018). The best-performing models of late in terms of internally validated prediction and forecasting have been members of the left path, *e.g.*, Yoon et al. (2016), and Rajkomar et al. (2018).

There are many important clinical tasks that would benefit from a tailored risk score–magnitudes more than the number of existing clinical studies producing high level evidence (Tricoci et al., 2009). At the same time, it is not clear if the results from algorithms that conduct multi-task learning produce clinically useful interpretations or actionable insights (*e.g.* micro- and macro F1 scores). In fact, for the multi-task problem posed in the work by Rajkomar et al. (2018), the authors elected to train separate models with distinct formulations for each task under consideration. We suggest the flexibility of machine learning models to conduct multi-task learning may be most useful where (1) the tasks are related to a single event but vary in other dimensions, for example forecast time, and (2) in addressing questions where the generalizability from older model studies may be suspect, for example generalizing from clean study populations to real-world populations with comorbidities. Therefore, we suggest targeted use of the multi-task formulation to better characterize individual disease processes.

Accordingly, our work introduces a forecasting method (on the left path in Figure 1) and explores a performance-simplicity-representation tradeoff in the multi-forecast setting. We introduce a marked point process based on wavelet reconstructions to predict targeted clinical events over time in patients with diabetes. We use wavelets to encode distributions over feature values and relative time and, like

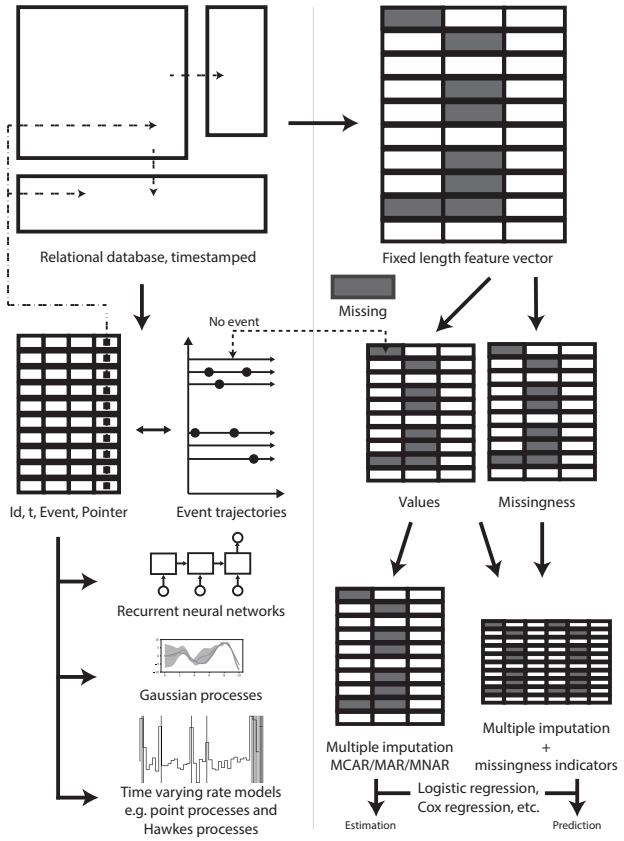

Figure 1. Log files, stored in relational databases, with extraction pipelines. Point process (left) and classical fixed-length feature vector (right) pipelines are contrasted.

multivariate Hawkes processes, map through reduction to a rate function on absolute time. We encode the wavelets, the relative-to-absolute mapping, and the reduction step together as a neural network.

To motivate the specific formulation, consider the limitations of the Hawkes process for clinical care processes. First, the Hawkes process encodes an additive relationship of change in rate from recurring precursors, *i.e.*, burstiness, whereas in health care, the repeated measurement of an event beyond the first, say of glucose, might be irrelevant. Second, clinical event timing may be routine, scheduled, or emergent, which suggests that kernel learning will improve model performance because changes in the rate may be time-dependent and not immediate. Third, clinical event processes are marked, with marks that could be categorical, real, or null values: *e.g.*, bacterial culture: staphylococcus aureus, glucose: 200, and ketoacidosis: NA.

To address the first limitation, summation, we encode a reduction layer, where we allow for reductions other than "sum" of kernel contributions from recurring events. To address the second and third limitations, non-specific timing and lack of marks, we propose a kernel learning method over

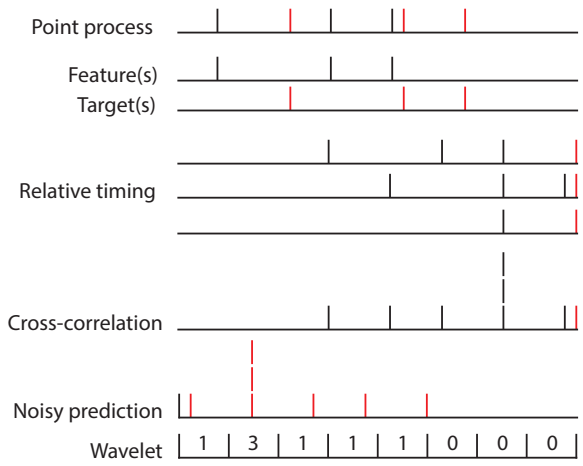

Figure 2. Illustrative 1-d cross-correlation motivating the discrete wavelet reconstruction kernel.

one dimension (time) and two dimensions (time-value) using wavelet reconstructions. The motivation for wavelets is illustrated in Figure 2, where a discrete wavelet reconstruction encodes the relationship of time-delayed events identified through cross-correlation. While cross-correlations capture all relative timings, many may be spurious or coincidental. The wavelet representation of relative timing instead is learned through likelihood optimization. To capture the effect of the value distribution of marks on events, two-dimensional wavelet reconstructions are value indexed, producing a one-dimensional reconstruction that is passed forward to the reduction layer.

With the introduction of reduction and kernel-learned timing and marks, we apply our model first in simulations of heart attacks and scheduled hemoglobin A1c (HgbA1c) checks. We then evaluate the model in real health records of patients at risk for diabetes. We show our model has strong performance and explore its interpretability characteristics in small data and over time.

1.1. Related work

Both neural networks and Hawkes process variants are used for survival analysis in health care: a few recent examples include Choi et al. (2015), Du et al. (2016), Alaa et al. (2017), and Bao et al. (2017). To effectively model the hazard two key properties are used: (1) the ability of Hawkes processes to capture relative timing of interdependent events, and (2) the flexible functional forms of neural networks that are able to capture relative timings, albeit somewhat opaquely. The closest work is likely that of Bao et al. (2017), where the authors adopt dyadic influence functions. However in that work marks are not used and the dyads selected are a subset of the Haar wavelet basis. Outside of health applications, related literature includes Hawkes kernel learning, *e.g.*

Zhou et al. (2013); Linderman & Adams (2014); Lee et al. (2016) and generalizing the Hawkes process with neural networks, *e.g.*, Mei & Eisner (2017). Our method follows these approaches, instead using a wavelet representation across value and time to capture long-range dependencies.

Our method of mapping relative timing hazard components onto absolute time possesses the advantages of the Hawkes approach and adopts a simple but well-performing neural network architecture. To achieve the mapping, a stepwise hazard approximation is made, as is done in Jing & Smola (2017) and Weiss (2017), however, instead of LSTMs and forests that have challenging interpretations, our method remains interpretable for small data sets, where exploration through visualization similar to that of Caruana et al. (2015) can be performed. Our method uses wavelets to compactly represent event contributions, and several survival analysis approaches have also adopted them in univariate models, *e.g.*, in Antoniadis et al. (1994) and Brillinger (1997). Outside of survival analysis and point processes, wavelet-inspired neural networks have seen success, with Wave-net adopting wavelets to classify time series (Bakshi & Stephanopoulos, 1993), and Wavenet adopting a multilayer hidden neural architecture to connect distant time steps (Van Den Oord et al., 2016).

1.2. Contributions

The contributions of this work are as follows: our work generalizes multivariate Hawkes processes to allow for non-additive event rate relationships. Like other works, *e.g.* Linderman & Adams (2014), our work learns the kernel function that relates multivariate event histories to the rate. However, in our work we use wavelets as the kernel, akin to a multivariate development from Brillinger (1997). We leverage the scaling property of wavelets to formulate a regularization that balances spatiotemporal generalizability with deterministic or near-deterministic event timing. Unlike many sequence models, *e.g.* Hochreiter & Schmidhuber (1997), which are affected strongly by choice of time step, our work adopts an absolute and relative time frame, and therefore the granularity of the absolute time domain need not be determined a priori. Additionally, unlike some point process formulations, our work models marks that are 1-d (event times) and 2-d (event times and their category or real value). Our work is explicit about types of forecasting tasks, and the methods are adapted for several purposes: forecast performance, interrogability, and representation. We show that our method performs competitively on several health care tasks and produces useful representations and predictions. Finally, we describe the hazard likelihood formulation and the time-varying contributions of events as a function of forecast distance as directions for work on multitask learning and neural network attention and attribution mechanisms.

2. Background

Let E be the set of events with target event $y \in E$ the event we want to forecast. Associated with each event e is a value $v \in V$ with $|E| = |V|$. An example consists of a sequence of (time, event, value) tuples and a period of interest for forecasting. For the n-th example, $n \in \{1 \dots N\}$, define T_n as the number of tuples. Then the sequence can be written as (t_{in}, e_{in}, v_{in}) for $i \in \{1 \dots T_n\}$, with the period of interest denoted as τ_{ny}.

Let $\lambda_y(t)$ be the rate functions of interest, dropping the subscript n for ease of notation. The multivariate Hawkes process can then be written as follows:

$$\lambda_y(t) = \lambda_0(t) + \sum_{e=1}^{|E|} \beta_e \sum_{i=1}^{T} g_e(t - t_i) \mathbb{1}(t_i < t, e_i = e)$$

where $\lambda_0(t)$ is a baseline population rate function, $g_e(\cdot)$ is a kernel function for event e relating its effect on the rate of y, β_e are event-specific parameters, and $\mathbb{1}(\cdot)$ is the indicator function. Typically $g_e(\cdot)$ is an event-specific exponential decay function with a learnable decay parameter. Self-exciting processes are defined by $g_y(\cdot) > 0$, bursty processes by $g_e(\cdot) > 0$, and inhibitory processes by $g_e(\cdot) < 0$. A few recent variations include Linderman & Adams (2014) who represent g_e with a Bayesian graph kernel and Xu et al. (2017) with the product of an infectivity function and a triggering kernel.

Given $\lambda_{ny}(t)$, the log likelihood of the data is:

$$\mathrm{LL}(X|\theta) = \sum_{n=1}^{N} \Big(\sum_{i=1}^{T_{ny}} \log \lambda_{ny}(t_{iny}) + \int_{\tau_{ny}} \lambda_{ny}(t) dt \Big) \tag{1}$$

The form of the Hawkes process is limiting, however, because (1) the effect of $g_e(\cdot)$ decays over time, (2) the effect over $g_e(\cdot)$ is additive, (3) the value associated with each event is not considered, and (4) the time restriction in the indicator function implies nowcasting ($\mathbb{1}(t_i < t)$) not forecasting ($\mathbb{1}(t_i < t - c)$ for some $c > 0$). Making modifications to achieve these characteristics is desirable to effectively model many real-world processes. For example, a patient with new-onset diabetes schedules an appointment with a typical gap interval of 3 months, and the presence of 1 or 10 elevated readings may not affect the timing of the scheduled appointment. The former suggests the utility of kernel learning, and the latter suggests summation over g_e is not the appropriate reduction.

The proposed method addresses these concerns. In particular, we adopt discrete wavelet reconstructions, which both allows kernel learning and the use of marks. Additionally, by representing the point process as a neural network, we are able to (1) use maximization alongside summation in

a reduction layer (the formulation enables specification of any number of reductions), and (2) conduct time-dependent censoring. We formalize the model below.

3. Method

Define J as the set of wavelet dimensions to be considered: we set $|J| = 2$, with $j = 1$ referring to time reconstructions, and $j = 2$ referring to time and event value reconstructions. Let w_{ej} be the wavelet coefficient tensors for event e and wavelet reconstruction dimension j. Define $\Phi = \{\phi_{ej} : (w_{ej}, v_e) \mapsto g_{es}\}$ as the set of wavelet reconstruction functionals with inputs wavelet coefficient tensors w_{ej} and event value v_e and output $g_{es} \colon t \mapsto \mathbb{R}$, where $s \in S$ is the active state of a discrete set S corresponding to event value v_e. Conceptually, given event value v_e, the wavelet reconstruction functional ϕ_{ej} reconstructs the signal from w_{ej} and indexes the space dimension(s) with v_e, producing function g_{es}, a function with inputs of time. Next, we introduce mapping $q = q(g(t), t_i) = g(t - t_i)$ that transforms function g_{es} on relative time to function g_d in absolute time, which is then passed to a set of reduce functions R. For our model, we will set $R = \{\text{sum}, \max\}$, though others could be considered. Finally, we incorporate time-dependent censoring with functional C. Let $C(c)\big(t, t_i\big)$ equal 1 if $t - t_i > c$ and 0 otherwise, and let $\psi_{ejc}(w_{ej}, v_e) = \psi_{ej}(c)(w_{ej}, v_e) = C(c) \circ \phi_{ej}(w_{ej}, v_e)$ where $\circ$ is the Hadamard product. Analogous to the Hawkes process model, we can write our model as follows:

$$\lambda_{yc}(t) = \lambda_0(t) + \sum_{e=1}^{|E|} \sum_{r,j}^{R,J} \beta_{erj} \times r_{i=\{1,\dots,T\}}^{\tau_y} \big(q(\psi_{ejc}(w_{ej}, v_i), t_i) \mathbb{1}(e_i = e)\big) \tag{2}$$

where $r_{i=\{1,\dots,T\}}^{\tau_y}$ indicates the reduction occurs over T functions over the interval τ_y. The parameters of the model are $\Theta = \{w_{ej}, \beta_{erj}\}$. Because the system may be overdetermined, we add regularization terms. The first is $\gamma_\beta \sum_e \sum_{rj} ||\beta_{erj}||_1$ akin to the LASSO (elastic-net regularization is equally straightforward). The second is the regularizer $\gamma_w \sum_{ej} ||u(w_{ej})||_1$ akin to sparse shrinkage on the wavelet tensor with a choice for u.

We define $u(w_{ej}) = \bigotimes_{k \in \{1,\dots,j\}} 2^{l_k/2} \circ w_{ej}$, where l_k is the wavelet scale parameter of the k-th dimension. The idea is that regularization on wavelets for point events corresponds to smoothing a function of Dirac deltas over time, and we want the log loss effect of a Dirac delta (an element of the first term in Equation 1) to be in proportion to the activation of the unnormalized wavelet basis function so that the data drive the choice of resolution. To do so, the regularization must be in proportion to $\bigotimes_{k \in \{1,\dots,j\}} 2^{l_k/2}$. An example is the orthonormal two-level Haar wavelet, where the orthonormal transformation matrix is written as the Hadamard product of exponentiated scale parameters and unnormalized basis functions:

$$\begin{bmatrix} 2^{-1} \\ 2^{-1} \\ 2^{-1/2} \\ 2^{-1/2} \end{bmatrix} \circ \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix}.$$

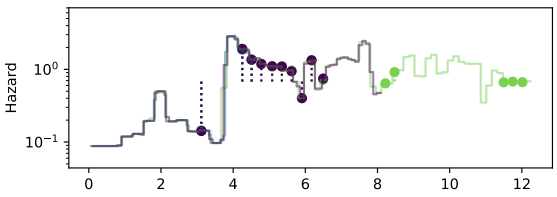

Figure 3. Three absolute time hazards (green, blue, and purple) for one trajectory with different steps sizes and censor times.

3.1. Relative- to absolute-time transformations

To map the wavelet reconstruction defined over relative time onto absolute time, we define causally-protective piecewise transformations. Let τ_d denote disjoint caglad intervals that comprise τ_y the target interval in absolute time, and let τ_e be the relative time intervals the wavelet reconstruction is defined over for event e. Then, for event time t_{in}, the absolute wavelet intervals are $\tau_{ine} = \tau_e + t_{in}$, and the transformation is given by $g_d = \tau_d^{-1} \sum_{\tau_{ine} \in \tau_d} g_{es} \tau_{ine}$. Because this transformation allows causal leakage in that a τ_{ine} may affect an interval of time in τ_d that both precedes and overlaps it, we apply a Hadamard censor to g_{es} before calculating g_d.

The convenient property of this formulation is that the granularity over absolute time can be adjusted with small effect on the hazard. Figure 3 illustrates an absolute-time hazard function of a single trajectory using different absolute time steps and applying censorship at different times, resulting in absolute hazards similar but not identical. Sharp discontinuities near target event times can result in relatively large likelihood differences, and the ability to choose the absolute time granularity τ_d post-training facilitates hazard recovery as needed. Figure 3 also acts as a causality leakage check by demonstrating that the hazard is unaffected by the presence or censorship of future events.

3.2. Censoring

Figure 4 illustrates the two dimensions of the forecasting task–the time at prediction and the desired forecasting time(s)–and four common forecasting procedures. Cox processes (Cox, 1972) forecast from time $t = 0$, the entry time of a study where baseline characteristics are measured, and rates are predicted for all future times t' with full censoring of time-varying features. "Anytime" forecasting denotes forecasting at any time t but without time-varying features. The censor distance $c > 0$ distinguishes nowcasting from

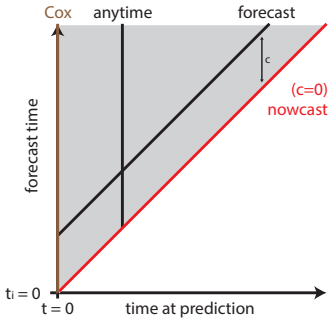

Figure 4. Forecasting tasks. The vertical distance from the line $t' = t$ is c the feature censoring distance. Common loss definitions are along solid lines.

forecasting ($c = 0$), and this is typical of clinical monitoring risks scores where new data is continuously arriving. The shaded gray area indicates the region of valid forecasts, *i.e.*, without causal leakage. Full use of the upper right quadrant occurs in post-hoc predictive tasks, and there is a literature on characterization after the fact, *e.g.*, using bidirectional RNNs (Graves & Schmidhuber, 2005).

Depending on the task, performance is evaluated on different lines of Figure 4. For nowcasting, where the next step prediction task is a forecast with a sawtooth censor dependent on the width of the time interval, we set $c = 0$. Forecasting is performed similarly with non-zero c. For recovering a representation of the underlying phenomenon, we introduce multi-forecasts: the weighted performance at multiple values of c. The idea is that the underlying relative-time representation should be invariant of the forecast censor c. To do so we expand a tensor dimension of c values with an indicator Hadamard censor applied to each row to censor the relative-time wavelet reconstructions. In this case, we optimize the performance of λ_{yc} for multiple values of c but constrain the model to share representation of w_{ej}.

3.3. Improving prediction

The formulation in Equation 2 can be seen as a generalized linear model of add-/max- reductions of wavelet reconstructions. While the generalized linear form lends itself to interpretation, we consider whether non-linearities will further improve predictive performance. We introduce permute-and-pool layers that randomly permute event ordering within time step, randomly select sign, perform max-pool, and project linearly to the next layer. In place of the double summation in Equation 2, we apply a random sign Hadamard Z and pass the result to P parallel permutation layers with max pools of size $\min(2^p, |E||J||R|)$ for $p = 0$ to $P - 1$. The outputs of the max pool are then linearly combined and output to the next layer (terminally the reduction layer).

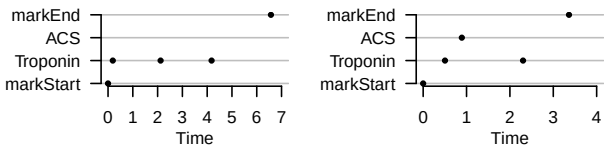

Figure 5. Acute coronary syndrome simulation, example trajectories of angina (left) and ACS (right).

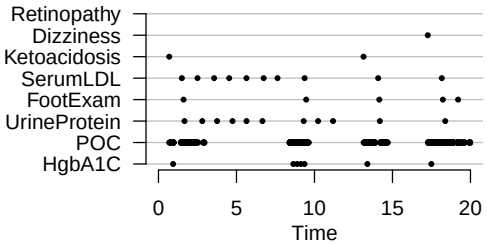

Figure 6. Diabetes simulation, example trajectory.

4. Experimental setup

We conduct tests on two simulated data sets of health care, and one real health care data set. We compare wavelet reconstruction networks (WRNs) with homogeneous Poisson processes, multivariate Hawkes processes, and long short-term memory (LSTM) networks. We compare against the algorithms in the nowcasting framework and evaluate using negative log likelihood. We then explore forecasting and multi-forecasting in WRNs.

For multivariate Hawkes process, we adopt a kernel with event-specific exponential decay parameter $\gamma_e > 0$: *i.e*, $g_e(t - t_i) = e^{-\gamma_e(t-t_i)}$. We use a learnable constant baseline rate λ_0. We learn β_e without constraint, rather than $\beta_e \geq 0$ or $\beta_e \leq 0$ to get Hawkes and inhibitory processes respectively. The comparison LSTM we implemented uses a linear input ($i \times h$), an LSTM hidden layer ($h \times h$), and a rectified linear output ($h \times 1$) architecture where h is the hidden unit width. For each comparison model, the output is a hazard per time step λ_{ny} that is fed into the survival log likelihood loss in Equation 1.

For each experiment, we divide the data into a train, tune, and held out test set. Model development is performed on train and tune sets. Once we are set on choice of parameter settings based on best early-stopping tune set log likelihood, we load the held out test set and apply the learned network. We compare the algorithms and report the average log likelihood on the held out test set.

There are number of settings to provide for the implementation. We set the number of absolute time steps to 80, equal in size. We treat categorical variables both as empty marks and as an indicator variable per category. We use one- and two-dimensional Haar wavelets with 64 relative time bins and 16 value bins. The time bins are linearly spaced from

Table 1. Descriptive statistics of the diabetes study population, reported with median [2.5%,97.5%] and *n* (fraction).

Feature	*n*=4,732
Age in 2010	35 [0, 82]
Gender	
Female	2526 (0.53)
Male	2204 (0.47)
HgbA1c	980 (0.21)
First	5.9 [4.8,10.2]
Any	6.7 [5.1,10.4]
Combined	137 (0.03)
Ketoacidosis	15 (0.00)
Polyneuropathy	88 (0.02)
Retinoptahy	34 (0.01)

0 to the largest relative difference of event and target in the training set. If a feature's train set value distribution is both positive and negative, the bins are mapped linearly across the value range, and if values are of one sign, bins are spaced linearly in the space of a $\log(1+|x|)$ feature transformation. In our experiments we use a single permute-pool layer with $P = 4$. We optimize using ADAM with manual setting of the learning rate $1e-3$, event/reduction regularization parameter $\gamma_\beta = 1e-7$, and wavelet regularization parameter $\gamma_w = 1/N$, all based on tune set performance. We use a mini-batch of size 10. For ease of computation, we randomly subsample features whose occurrences exceeded 50 to 50. Inspection of the data suggests that the subsampling would have negligible effect on performance while reducing memory requirements substantially. We penalize hazards below $1e-5$ by a large constant 100 per unit time on the train set, and pass all predictions through a rectified linear unit to ensure valid hazards. We set a max hazard gain of $\log(10)$ to ensure numeric stability. These constraints are appropriate for our tasks, as rates below $1e-5$ and above 10 are not meaningful for our applications. For LSTMs, we set $h = 4$ and 16 for the simulated and real data experiments.

4.1. Simulations

The first simulation is of heart attack diagnoses denoted by acute coronary syndrome (ACS). In this simulation, it is the elevation in value of troponin, a heart enzyme measurement, outside the normal range (less than 0.01 ng/mL) that indicates ACS will occur in the next time unit uniformly at random. Figure 5 illustrates a trajectory; note both the mark and the timing are important for ACS determination. The second simulation is of diabetes care: patients with diabetes undergo semi-regular appointments, *e.g.*, annual eye and foot exams, quarterly hemoglobin A1c measurements, and pre- and post-prandial glucose measurements. These patients are often non-adherent with worsening adherence as a function of increasing time from adverse events: dizziness, ketoacidosis, and retinopathy. Figure 6 illustrates the timings of an example trajectory.

Table 2. Negative log likelihood by algorithm and data set for nowcasting on the test set. Lower number is better. * denotes simulation; KNR: {ketoacidosis, neuropathy, retinopathy}; H. Poission: homogeneous Poisson; WRN: wavelet reconstruction network; PPL: with permute pool layer. Best performer in bold.

Method	ACS*	A1c*	A1c	KNR
H. Poisson	0.44	18.54	2.86	0.75
Hawkes	0.39	3.87	1.67	0.31
LSTM	**0.13**	4.10	1.29	0.35
WRN	0.23	3.93	1.23	**0.24**
WRN-PPL	0.15	**3.78**	**1.13**	0.26

4.2. Diabetes visits

We partnered with a regional health care system to investigate the risk of adverse outcomes of diabetes and the care those patients received. From the population included in the regional cohort, followed from 2010 to 2017, we selected those at risk of diabetes as defined by an outpatient measurement of hemoglobin A1c or glucose, or a diagnosis of hyperglycemia. Among those, we excluded any individuals without at least two clinic encounters more than six months apart. We additionally applied a censor date at the time of the last clinical event before a 30-month gap in care, where there is substantial uncertainty that the patient is lost to follow-up or is receiving care outside of network.

Application of the inclusion and exclusion criteria resulted in 798,818 timestamped events in a study population of 4,732 individuals. We divided the population into thirds: {train, tune, test} sets. We focused on two outcomes: (1) hemoglobin A1c measurements, as a proxy for scheduled diabetes care, and (2) a combined outcome of {ketoacidosis, neuropathy, retinopathy} as defined by ICD 9 and ICD 10 codes. Features included were extracted with string matching on event descriptions of events documented as putative risk factors in clinical guidelines from the ADA, AHA, and UpToDate, and included events from demographics, medications, encounters, laboratory, diagnosis, and procedures tables. The extraction resulted in 575 features. Dates of events were perturbed across years and subsampled. Table 1 provides descriptive statistics on the study population and outcomes. Train and tune set loss-by-epoch curves are provided in the Appendix.

5. Results

Table 2 reports the negative log likelihoods for the experiments on the held out test sets. Overall, the proposed wavelet reconstruction networks outperformed the other algorithms. The WRN-PPL method excelled most on tasks with many target occurrences (A1c* and A1c experiments) and performed near to the best in rare occurrence data (ACS* and KNR). The Hawkes process performed less well

with comparable performance only in the A1c* data set. The LSTM performed well across all data sets when given enough random restarts during training (approximately 10 restarts). The WRN method outperformed the WRN-PPL method at the KNR task, which could be due to the relatively low rate of target events (0.025 per year) where complexity may lead to overfitting; however the difference in negative log likelihood is small. The effect of WRN-PPL forecast distance c on KNR prediction is shown in the Appendix. Notably, the 3-month censored WRN-PPL has the same performance as the nowcasting LSTM.

Figure 7 shows the hemoglobin A1c predicted hazards profile for two random test set patients using WRN-PPL and one month forecasts. The step function represents predicted hazard over time, points indicate true event times and dotted line show the difference from the baseline (homogeneous Poisson) rate. The WRN-PPL algorithm makes predictions that anticipate appointments where hemoglobin A1c will be measured in quasi-periodic fashion. From the prediction, it appears there are missed occurrences, which could be informative for scheduling, adherence, or reminder systems. Similarly, Figure 8 illustrates the ability to model the rates of adverse outcomes. Medical guidelines do not specify scheduling for regular follow-up of the adverse events, and this is congruent with the lack of periodicity in the KNR hazard predictions. Additional random test set hazard figures are available in the Appendix.

Figure 9 shows the wavelet reconstruction for the effect of troponin level and timing on the hazard. Both reconstructions demonstrate recovery that acute coronary syndrome is diagnosed within the next time unit of a troponin above 0.01 ng/mL. The multiple-c reconstruction on the right more accurately reflects the hazard of a uniform distribution, with increasing hazard if the event has not yet occurred.

Additional effects of multiple-c prediction are shown in Figure 10. The left plot indicates how early in advance the outcome can be predicted. Figure 10 also shows that the WRN does not have adequate flexibility to effectively model large and small c predictive tasks effectively. The WRN-PPL method shows improvement, indicated by the sharp improvement at $c = 1$, however, the performance is not as strong as when predicting at a single value of c, as indicated in Table 2. The coefficient profile as a function of c in multiple-c prediction is also shown, demonstrating that relative-time attributions are censor-dependent.

6. Discussion

The performance of WRN-PPL in Table 2, Figure 7, and Figure 8 illustrates the effectiveness of our model, in particular in identifying recurring events. The peaks reach hazards of approximately 3, indicative of a mixture of belief and uncertainty–belief that in those periods the event should occur at a rate above three per year, and uncertainty about the occurrence or precision of the appointment.

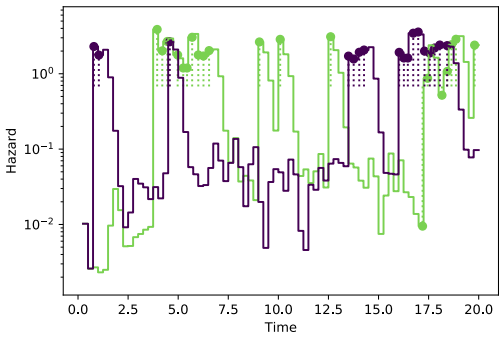

Figure 7. Hazards and hemoglobin A1c* events for two random patients in the test set.

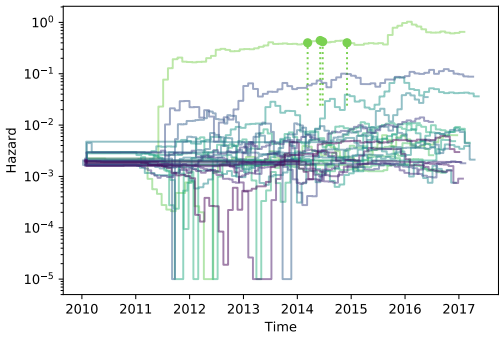

Figure 8. Hazards and ketoacidosis, retinopathy, or polyneuropathy diagnoses for thirty random patients in the test set.

We also attempt to quantify the improvement in negative log likelihood. From Table 2, on average the WRN prediction is $e^{2.86-1.13} = e^{1.73} = 5.6$ times more likely than the baseline hazard (homogeneous Poisson) in predicting hemoglobin A1c tests. Caution is warranted in this interpretation because the estimate at a specific time for a specific individual may still be unreliable. For example, predicting a hazard of 0.01 when the true hazard is 1 invokes a penalty of $log(1) - log(0.01) = 4.6$ when an event occurs and a per time unit benefit of $1 - 0.01 \approx 1$ when no event occurs. If on average one event occurs per time unit, the suboptimality of the log likelihood is ≈ 3.6. However, if this occurs once in the data set of size 1578, then the change in average log likelihood is 0.002, which is one to two orders of magnitude smaller than the differences in log likelihood reported in Table 2. Unfortunately, because of disproportionate attention given to predictions of large hazards, as described in the next paragraph, translation of the delta log likelihood into an quantifiable statement is challenging.

While use of the negative log likelihood demonstrates effective hemoglobin A1c modeling in Figure 7, also shown in

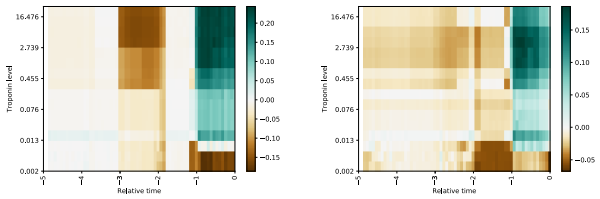

Figure 9. Haar wavelet reconstruction of troponin contribution to ACS hazard. WRN-PPL reconstruction image for nowcasting (left) and multi-forecasting at $c = \{0, 0.25, 0.5, 0.75, 1\}$ (right). Best viewed in color.

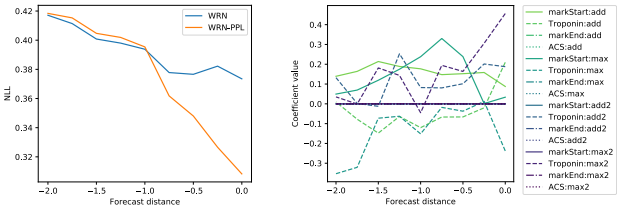

Figure 10. Negative log likelihood as a function of forecast censoring distance c for multi-forecasts (left). The permute and pool layer expresses greater flexibility to model the hazard than WRN, but is not adequately flexible to perform as well as single-c WRN-PPL at each censor distance. Coefficient profile as a function of forecast censoring distance c for WRN (right). The reduction layer comprises additions and maximums of 1-d and row-indexed 2-d reconstructions.

the Figure is the failure to identify the patient represented by green as high risk at the time of one of the hemoglobin A1c measurements. We suggest that it is through the optimization of negative log likelihood that this occurs. The problem is that individuals with frequent target events dominate the optimization because errors hazard prediction are more costly. To illustrate this, compare the cost of predicting a hazard of 2 when the hazard is 1 versus predicting a hazard of 0.2 when the hazard is 0.1, both over the period 2010 to 2017. The former invokes on average a penalty of $7 - 7\log 2 \approx 2.14$, while the latter invokes a penalty of $0.7 - 0.7(\log 0.2 - \log 0.1) \approx 0.214$. Therefore, if an algorithm is unable to model both the high hazard and low hazard proportionally well, it optimizes for the high hazard at the cost of the low hazard. Nevertheless, in survival analysis the hazard likelihood is the most commonly used objective function. Classic solutions to this problem involve manual attention mechanisms, for example, censoring the data at the time of first target event.

A comparison of the results in Figure 10 and Table 2 demonstrates the value of flexibility of the model in multi-forecast learning. In particular, Table 2 shows that single forecasting outperforms multi-forecasting at $c = 0$ in Figure 10. However, Figure 9 illustrates that multi-forecasting improves the learned wavelet representation. These findings suggest that the layering between the wavelet reconstruction–{reduction layer, linear} (WRN) and {reduction, permute and pool, linear} (WRN-PPL)–and the hazard output is not adequately flexible to map the true wavelet reconstruction to the true hazard. We argue the solution is not in simplification or abandonment of the multi-task setting, but in leveraging the multi-task setting to facilitate recovery of the wavelet reconstruction by using an even more flexible mapping from reconstruction to hazard.

Figure 10 (right) illustrates another challenge of using temporal associations to make statements about the association of features on the outcome. Specifically, the coefficient estimates vary over censor time c. For example, the coefficient for a feature that is noisily related to the outcome, *e.g.*, a distant cause, may be fully ignored due to an intermediate variable, *e.g.* a proximal cause, for small c but not for large c. Thus, putative distant causes and intermediate variables could be reflected in the shifts in coefficient profile as a function of c, and the strength of cause could be related to the degradation in predictive performance for increasing values of c. That said, our experiments were conducted on observational data, and therefore investigation of the coefficient profile for putative causes serves primarily for hypothesis generation. We refer the reader to recent works in machine learning that model the counterfactual distribution through the potential outcomes framework, *e.g.*, that of Johansson et al. (2016), Schulam & Saria (2017), and Yoon et al. (2018) and to literature on testing new policies with randomized controlled trials (Friedman et al., 2015).

7. Conclusion

We formulated a multivariate Hawkes-like process using neural networks enabled reduction of recurring events, kernel learning, and use of marks. We adopted a Haar wavelet representation of time and time-value to encode the relationships between features and the outcome. We demonstrated improved performance of our system over Hawkes processes and LSTMs in patients with diabetes. We showed the ability to capture quasi-periodic events that could be used to measure adherence. Our results also lead us to make observations about the implicit attention mechanism of likelihood to high risk individuals, about tradeoffs of flexibility for representation and performance, and about varying time-varying contributions as a function of forecast distance.

While our work demonstrates an effective framework for forecasting from marked time-stamped data with choice of forecast timing, multi-forecasting, reduction and permute-pool layering depending on the use case, there are many avenue of future investigation. One avenue will be to formulate methods that performance and representation simultaneously. Another avenue will be the development of forecast-

ing methods that leverage embeddings of marked features. Greater attention to the attention bias in the hazard likelihood is warranted so that level sets of likelihood are spread across individuals and time. This characteristic of the hazard likelihood will also be an important consideration for future work in the multi-forecast setting, where implicitly shorter forecasts receive more attention due to their ability to better model high hazards.

References

Alaa, Ahmed M, Hu, Scott, and van der Schaar, Mihaela. Learning from clinical judgments: Semi-markov-modulated marked Hawkes processes for risk prognosis. *arXiv preprint arXiv:1705.05267*, 2017.

Antoniadis, A, Gregoire, G, and McKeague, IW. Wavelet methods for curve estimation. *Journal of the American Statistical Association*, 89(428):1340–1353, 1994.

Asch, David A, Troxel, Andrea B, Stewart, Walter F, Sequist, Thomas D, Jones, James B, Hirsch, AnneMarie G, Hoffer, Karen, Zhu, Jingsan, Wang, Wenli, Hodlofski, Amanda, et al. Effect of financial incentives to physicians, patients, or both on lipid levels: a randomized clinical trial. *Jama*, 314(18):1926–1935, 2015.

Bakshi, Bhavik R and Stephanopoulos, George. Wavenet: A multiresolution, hierarchical neural network with localized learning. *AIChE Journal*, 39(1):57–81, 1993.

Bao, Yujia, Kuang, Zhaobin, Peissig, Peggy, Page, David, and Willett, Rebecca. Hawkes process modeling of adverse drug reactions with longitudinal observational data. In *Machine Learning for Healthcare Conference*, pp. 177–190, 2017.

Brillinger, David R. Some wavelet analyses of point process data. In *Signals, Systems & Computers, 1997. Conference Record of the Thirty-First Asilomar Conference on*, volume 2, pp. 1087–1091. IEEE, 1997.

Caruana, Rich, Lou, Yin, Gehrke, Johannes, Koch, Paul, Sturm, Marc, and Elhadad, Noemie. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1721–1730. ACM, 2015.

Centor, Robert M, Witherspoon, John M, Dalton, Harry P, Brody, Charles E, and Link, Kurt. The diagnosis of strep throat in adults in the emergency room. *Medical Decision Making*, 1(3):239–246, 1981.

Choi, Edward, Du, Nan, Chen, Robert, Song, Le, and Sun, Jimeng. Constructing disease network and temporal progression model via context-sensitive Hawkes process. In *Data Mining (ICDM), 2015 IEEE International Conference on*, pp. 721–726. IEEE, 2015.

Cox, David. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):87–22, 1972.

D'Agostino, Ralph B, Vasan, Ramachandran S, Pencina, Michael J, Wolf, Philip A, Cobain, Mark, Massaro, Joseph M, and Kannel, William B. General cardiovascular risk profile for use in primary care: the framingham heart study. *Circulation*, 117(6):743–753, 2008.

Du, Nan, Dai, Hanjun, Trivedi, Rakshit, Upadhyay, Utkarsh, Gomez-Rodriguez, Manuel, and Song, Le. Recurrent marked temporal point processes: Embedding event history to vector. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1555–1564. ACM, 2016.

Friedman, Lawrence M, Furberg, Curt, DeMets, David L, Reboussin, David, and Granger, Christopher B. *Fundamentals of clinical trials*. Springer, 2015.

Futoma, Joseph, Hariharan, Sanjay, and Heller, Katherine. Learning to detect sepsis with a multitask Gaussian process RNN classifier. In *International Conference on Machine Learning*, pp. 1174–1182, 2017.

Gardner-Thorpe, J, Love, N, Wrightson, J, Walsh, S, and Keeling, N. The value of modified early warning score (mews) in surgical in-patients: a prospective observational study. *The Annals of The Royal College of Surgeons of England*, 88(6):571–575, 2006.

Graves, Alex and Schmidhuber, Jürgen. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, 18(5): 602–610, 2005.

Hawkes, Alan G. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58(1):83–90, 1971.

Hochreiter, Sepp and Schmidhuber, Jürgen. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Jing, How and Smola, Alexander J. Neural survival recommender. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, pp. 515–524. ACM, 2017.

Johansson, Fredrik, Shalit, Uri, and Sontag, David. Learning representations for counterfactual inference. In *International Conference on Machine Learning*, pp. 3020–3029, 2016.

Lee, Young, Lim, Kar Wai, and Ong, Cheng Soon. Hawkes processes with stochastic excitations. In *International Conference on Machine Learning*, pp. 79–88, 2016.

Linderman, Scott and Adams, Ryan. Discovering latent network structure in point process data. In *International Conference on Machine Learning*, pp. 1413–1421, 2014.

Lipton, Zachary C, Kale, David C, Elkan, Charles, and Wetzel, Randall. Learning to diagnose with lstm recurrent neural networks. *arXiv preprint arXiv:1511.03677*, 2015.

Mei, Hongyuan and Eisner, Jason M. The neural Hawkes process: A neurally self-modulating multivariate point process. In *Advances in Neural Information Processing Systems*, pp. 6757–6767, 2017.

Rajkomar, Alvin, Oren, Eyal, Chen, Kai, Dai, Andrew M, Hajaj, Nissan, Liu, Peter J, Liu, Xiaobing, Sun, Mimi, Sundberg, Patrik, Yee, Hector, et al. Scalable and accurate deep learning for electronic health records. *arXiv preprint arXiv:1801.07860*, 2018.

Schulam, Peter and Saria, Suchi. Counterfactual Gaussian processes for reliable decision-making and what-if reasoning. In *Advances in Neural Information Processing Systems*, pp. 1696–1706, 2017.

Seymour, Christopher W, Kahn, Jeremy M, Martin-Gill, Christian, Callaway, Clifton W, Yealy, Donald M, Scales, Damon, and Angus, Derek C. Delays from first medical contact to antibiotic administration for sepsis. *Critical care medicine*, 45(5):759–765, 2017.

Tricoci, Pierluigi, Allen, Joseph M, Kramer, Judith M, Califf, Robert M, and Smith, Sidney C. Scientific evidence underlying the acc/aha clinical practice guidelines. *Jama*, 301(8):831–841, 2009.

Van Buuren, Stef and Oudshoorn, Karin. Flexible multivariate imputation by mice. *Leiden, The Netherlands: TNO Prevention Center*, 1999.

Van Den Oord, Aaron, Dieleman, Sander, Zen, Heiga, Simonyan, Karen, Vinyals, Oriol, Graves, Alex, Kalchbrenner, Nal, Senior, Andrew, and Kavukcuoglu, Koray. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

Weiss, Jeremy C. Piecewise-constant parametric approximations for survival learning. In *Machine Learning for Healthcare Conference*, pp. 1–12, 2017.

Xu, Hongteng and Zha, Hongyuan. A Dirichlet mixture model of Hawkes processes for event sequence clustering. In *Advances in Neural Information Processing Systems*, pp. 1354–1363, 2017.

Xu, Hongteng, Luo, Dixin, and Zha, Hongyuan. Learning Hawkes processes from short doubly-censored event sequences. In *International Conference on Machine Learning*, pp. 3831–3840, 2017.

Yoon, Jinsung, Alaa, Ahmed, Hu, Scott, and Schaar, Mihaela. Forecasticu: a prognostic decision support system for timely prediction of intensive care unit admission. In *International Conference on Machine Learning*, pp. 1680–1689, 2016.

Yoon, Jinsung, Jordan, James, and van der Schaar, Mihaela. GANITE: Estimation of individualized treatment effects using generative adversarial nets. In *International Conference on Learning Representations*, 2018.

Zhou, Ke, Zha, Hongyuan, and Song, Le. Learning triggering kernels for multi-dimensional Hawkes processes. In *International Conference on Machine Learning*, pp. 1301–1309, 2013.