Wavelet Reconstruction Networks for Marked Point Processes

Jeremy C. Weiss
JEREMYWEISS@CMU.EDU
Carnegie Mellon University, Heinz College of Information Systems and Public Policy

Abstract

Timestamped sequences of events, pervasive in domains with data logs, *e.g.*, health records, are often modeled as point processes or 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. Methods in point processes and recurrent neural networks capably model rate functions but their complexity may make interpretation, use and reuse challenging. Our work develops a high-performing and interrogable yet simple 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 these simple models achieve improved performance when applied to forecasting complications and care visits in patients with diabetes.

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). When used for early warning, risk scores have been associated with reduced mortality (Seymour et al., 2017). Underlying these approaches is the formulation of risk over time given some set of features. For example, Cox and Hawkes models make assumptions of proportional hazards and summation over kernel activations, which are often inappropriate in health settings.

In particular, problems arise when the repeated measurement of an event beyond the first, say of glucose, might be irrelevant. In addition, clinical event timing may be routine, scheduled, or emergent, which suggests that kernel learning will improve model performance because rate changes may be delayed and not immediate. Finally, clinical event processes are marked and possess variable types, *e.g.*, bacterial culture (Staphylococcus Aureus), glucose (200 mg/dL), and ketoacidosis. Multi-dimensional wavelets on relative time can address these issues, where a discrete wavelet reconstruction encodes the relationship of time-delayed events identified through cross-correlation (see Figure 1 for the 1-d case). We show that we can encode the wavelets, a relative-to-absolute mapping, and the reduction step on a computation graph to conduct learning. We apply our model first in simulations and in a diabetes application.

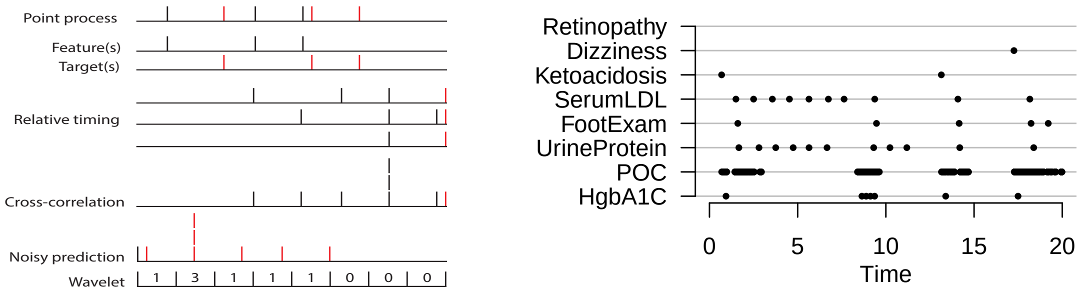

Figure 1: Illustrative 1-d cross-correlation motivating the discrete wavelet reconstruction kernel for relative time dependencies between one feature and outcome (left). Example: A1c* diabetes simulation (right).

Related work. Both neural networks and Hawkes process variants are used for rate modeling in health care: a few examples include Choi et al. (2015), Du et al. (2016), Alaa et al. (2017), and Weiss (2017). 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. Compared to Mei and Eisner (2017) that generalizes the Hawkes process with neural networks, our method (1) allows events to have marks, (2) enables forecasting and multi-forecasting (not just nowcasting), and (3) has a generalized linear form.

Our method uses wavelets to 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 using wavelets to classify time series (Bakshi and Stephanopoulos, 1993), and Wavenet adopting a multi-layer hidden neural architecture to connect distant time steps (Van Den Oord et al., 2016).

Contributions. The contributions of this work are as follows: first, we contribute a model that generalizes multivariate Hawkes processes to allow for non-additive event rate relationships. Like other works, our approach involves learning the kernel function that relates multivariate event histories to the rate. However, in our work we use wavelets as the kernel and can be seen as 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, 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 frame need not be determined a priori. We show that our method outperforms comparison algorithms in predicting complications and forecasting adherence. Finally, we show that our method characterizes adherence and risk of complications in an important health application.

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 $v \in V$. 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 variations include Linderman and Adams (2014) where g_e is a Bayesian graph kernel and Xu et al. (2017) where g_e is an infectivity function and triggering kernel product.

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 $c > 0$).

Wavelet reconstruction networks

We now define wavelet reconstruction networks (WRNs). We specify the form of the rate function, define our kernel function, and impose restrictions on the kernel function for forecasting and multi-forecasting.

Let $h_{ej}(t; t_i, \tau_d)$ be a piecewise-constant kernel function for event e on absolute time intervals τ_d with index set $J = \{1, 2\}$, where $j = 1$ indicates a time kernel and $j = 2$ indicate a time-value kernel. Let R be a set of reduce functions, which for our model we set $R = \{\text{sum}, \max\}$. Then $r^{\tau_y}_{i=\{1,\dots,T\}}$ indicates the reduction over T functions over the interval τ_y. We propose the following rate function:

$$\lambda_y(t) = \lambda_0(t) \; + \sum_{e=1}^{|E|} \sum_{r,j}^{R,J} \beta_{erj} r^{\tau_y}_{i=\{1,\dots,T\}} \big(h_{ej}(t; t_i, \tau_d) \mathbb{1}(e_i = e) \big). \tag{1}$$

Note that the kernel function h_{ej} is on absolute time, whereas g_e is defined on relative time in the previous section. This is done to specify the translation of the wavelet reconstructions on discrete relative-time intervals onto discrete absolute-time intervals, which requires additional treatment to prevent causal leakage. We define the kernel function on relative time below, followed by the definition of h_{ej} using the translation function.

Discrete wavelet reconstruction kernel. Recall that the discrete wavelet transform (DWT) is an invertible transform of a signal between time space and time-frequency space used in multi-resolution analysis and signal compression (Mallat, 1989). Here we encode our parameters in the time-frequency space and use the inverse DWT to reconstruct the relative-time kernel functions as follows.

For each event, we use one- to two- dimensional wavelets, with $j = 1$ referring to time reconstructions, and $j = 2$ referring to time and event value reconstructions.

We use discrete wavelets of size (α) and $(\alpha, |S|)$, with α time intervals on the interval $[0, \max_n(\max_{t_{iny}} t_{iny} - \min_{t_{ine}} t_{ine})]$ on the time dimension, and with $s \in S$ disjoint value intervals in $[min(v), max(v)]$ on the value dimension . For point events, we use the notation $s \in S$ where $|S| = 1$, and categorical events are treated as separate point events. Note that $j = 1$ reconstructions are the temporal analogues of missingness indicators.

Let parameters w_{ej} be the wavelet coefficient tensors for event e and wavelet reconstruction dimension j, and let g_{ejs} be the kernel function on relative time for event e and interval s, with $g_{ejs}(t - t_i) = 0$ for $t - t_i < 0$. Define the set of wavelet reconstruction functionals by: $\Phi = \{\phi_{ej} : (w_{ej}, v_e) \mapsto g_{ejs}\}$. Conceptually, given event value v_e, the wavelet reconstruction functional ϕ_{ej} reconstructs the signal from w_{ej} and indexes the value dimension with v_e, producing function g_{ejs}, a function with inputs of relative time.

Relative- to absolute-time transformations. To relate the absolute-time kernel h_{ej} with the relative-time discrete kernel g_{ejs}, we define the following causally-protective translation function. 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 of the wavelet reconstruction. We denote lower and upper endpoints with $\lfloor\cdot\rfloor$ and $\lceil\cdot\rceil$. Then, for event time t_i, the absolute wavelet reconstruction intervals have endpoints $\lfloor\tau_{ie}\rfloor = \lfloor\tau_e\rfloor + t_i$ and $\lceil\tau_{ie}\rceil = \lceil\tau_e\rceil + t_i$. The transformation q is given by:

$$h_{ej}(t; t_i, \tau_d) = q\big(g_{ejs}(t - t_i), \tau_d\big) = \frac{\sum_{\tau_{ie}\wedge\tau_d \not\emptyset, \lfloor\tau_{ie}\rfloor \geq \lceil\tau_d\rceil} (\max(\lfloor\tau_d\rfloor, \lceil\tau_{ie}\rceil) - \lfloor\tau_{ie}\rfloor) g_{ejs}(t - t_i)}{\lceil\tau_d\rceil - \lfloor\tau_d\rfloor},$$

where $\wedge$ denotes interval intersection. The second condition for inclusion in summation, $\lfloor\tau_{ie}\rfloor \geq \lceil\tau_d\rceil$, prevents causal leakage by ignoring intervals τ_{ie} that affect an interval τ_d that both precedes and intersects it. An advantage of our relative-time specification over RNNs (and their analogues) is that the granularity over absolute time can be adjusted with minimal effect on the hazard. By comparison, RNNs need to be retrained or reformulated if there are changes to the time-step specification, and RNNs that encode the inter-arrival times as features are not robust to irrelevant injected events.

Forecasting. Thus far we have specified h_{ej} as a nowcasting kernel because g_{ejs} is zero and non-contributory when $t - t_i < 0$ for all e and s. For forecasting, we incorporate time-dependent censoring with functional C, the Hadamard censor, and hyperparameter c the censoring distance, to prevent recent events, *i.e.*, risk modifiers, from affecting the rate. 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. We can specify the forecasting rate function analogous to Equation 1 as follows:

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

For multi-forecasting, we choose a vector of desired relative forecast times $\{c\}$ and maximize the average log-likelihood over all c. This may be distinguished from training separate forecasting models because the parameters of the model are tied.

Learning. 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, where we define $u(w_{ej}) = \bigotimes_{k\in\{1,\ldots,j\}} 2^{l_k/2} \circ w_{ej}$, where l_k is the wavelet scale parameter of the k-th dimension.

Table 1: Negative log likelihood, asterisk (*) denotes simulation.

Dataset	Method (NLL)							
	H. Poisson	Time-invarant	Nowcast	Hawkes	LSTM^1	LSTM^2	WRN	WRN-PPL
ACS*	0.44	0.43	0.36	0.39	0.21	**0.13**	0.23	0.15
A1c*	18.54	19.20	13.56	3.87	11.80	4.10	3.93	**3.78**
A1c	2.86	2.76	2.52	1.67	1.15	1.29	1.23	**1.13**
KNR	0.75	0.71	0.58	0.31	0.46	0.35	**0.24**	0.26

Improving prediction. While the generalized linear form (Equation 2) lends itself to interpretation, we consider whether non-linearities will further improve predictive performance. Therefore we introduce permute-and-pool layers (WRN-PPL) 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 ($\{-1, 1\}$) Hadamard tensor 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.

Experiments

We conduct tests in two simulations and on a health care data set. For direct comparison, we evaluate performance against methods in the nowcasting framework using negative log likelihood and ranking measures. We divide the data into a train, tune, and held out test set. Model development is performed on train and tune sets with parameters determined by early stopping. Models are then evaluated on the held out test set. We use the Goodman-Kruskal γ statistic as a measure of concordance among non-tied pairs. Further details of parameter setting, *e.g.* preprocessing, bin widths, optimizer settings, are in the Appendix. We compare wavelet reconstruction networks (WRNs) with homogeneous Poisson processes, time-invariant and nowcasting Fourier basis functions, multivariate Hawkes processes, and two long short-term memory (LSTM) networks. Details are given in the Appendix.

Data. 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. 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 (Figure 1, right). **EHR data:** We also partnered with a regional health care system to investigate the risk of adverse outcomes of diabetes and adherence to the care those patients received. Preprocessing led a study population of 4,732 individuals monitored over 7 years evenly split into training, validation and test sets. Details of the cohort are provided in the Appendix. 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 demographics and diabetes-related EHR features.

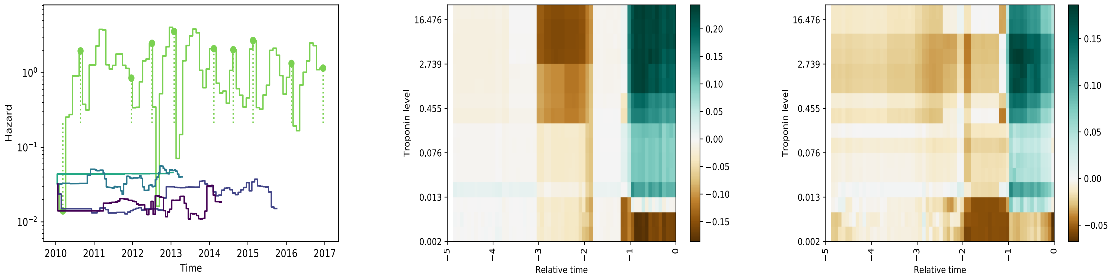

Figure 2: Left: Predicted rates of hemoglobin A1c tests for five random, test set patients (one month forecasts). Actual events correspond to circles, and dotted lines begin at the base risk and end at the predicted risk. Middle and right: Wavelet reconstruction of troponin contribution to ACS hazard, nowcasting and multi-forecasting. A preceding troponin above 0.01 indicates increased rate of ACS occurrence within the next hour (blue region).

Results

Overall, the proposed wavelet reconstruction network WRN-PPL outperformed the other algorithms (Table 1). The WRN-PPL method excelled in high-rate tasks (A1c* and A1c experiments) and were comparable to the best in low-rate tasks (ACS* and KNR). The WRN and WRN-PPL γ statistics showed effective risk stratification with a γ of 0.98, and concordance levels were comparable between the WRN and LSTM models (Table A1).

The WRN-PPL algorithm makes predictions that anticipate appointments where hemoglobin A1c will be measured in quasi-periodic fashion, while keeping the hazard low in-between (Figure 2, left). The model learns the effect of troponin level and timing on heart attack rate (Figure 2 middle, right). Both reconstructions demonstrate recovery that acute coronary syndrome is diagnosed within the next time unit after a troponin greater than 0.01 ng/mL. The multiple-c reconstruction on the right more accurately reflects the uniform distribution hazard, namely, increasing hazard if the event has not yet occurred.

The performance of WRN-PPL in Table 1 and Figure 2 illustrates the utility of our model, in particular in identifying the near-periodicity of recurring events. For example, the rate prediction for the individual denoted in green in Figure 2 (right) suggests that individual may have skipped, missed, or rescheduled 5 to 6 appointments over the last decade. The peaks reach hazards of approximately 3, indicative of a mixture of belief and uncertainty–belief that in those months the event should occur at a rate above three per year, and uncertainty about the occurrence of the appointment.

Discussion. Wavelet reconstruction networks is a simple forecasting method that performs competitively with neural networks. Its advantages include multi-resolution representation of relative time dependencies in 1 and 2 dimensions and enables time step specification at test time. We demonstrated the strong performance of our system in a analysis of diabetes and showed the ability to capture quasi-periodic events that could be used to measure adherence and forecast risk of complications. Future work will consider the use of more expressive wavelet families, and how to automatically detect time gaps (c) where performance gains are large, suggestive of time windows for early warning.

References

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

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

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

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

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

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

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

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

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

Nan Du, Hanjun Dai, Rakshit Trivedi, Utkarsh Upadhyay, Manuel Gomez-Rodriguez, and Le Song. 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*, pages 1555–1564. ACM, 2016.

J Gardner-Thorpe, N Love, J Wrightson, S Walsh, and N Keeling. 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.

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

Zachary C Lipton, David C Kale, and Randall Wetzel. Modeling missing data in clinical time series with rnns. *Machine Learning for Healthcare*, 2016.

Stephane G Mallat. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE transactions on pattern analysis and machine intelligence*, 11(7):674–693, 1989.

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

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

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

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

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

Appendix

Experimental setup details

The Fourier comparison methods are given by $f(t|t_0, x) = \sum_k \sum_l w_{kl} sin((2\pi l/\tau)(t - t_0)) + v_{kl} cos((2\pi l/\tau)(t - t_0))x_k$, where x corresponds to features, *i.e.*, event by value interval occurrences, at or before time 0 ($t_0 = 0$, time-invariant) and at $t_0 = t$ (nowcasting), k indexes the event by value interval by time step, and l indexes the basis function component. L2 regularization are applied to w_{kl} and v_{kl}. Then the rate function is defined as $\lambda(\cdot) = w_0 + f(\cdot)^2$. For nowcasting, the Fourier method is given features from 16 previous time steps, and since in nowcasting $t - t_0 = 0$, the formula reduces to a generalized linear model of $v_{kl}x_k$ terms. The multivariate Hawkes process we use includes 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 $\beta_e \leq 0$ of Hawkes and inhibitory processes respectively. We impose a positivity constraint to ensure the rate is non-negative.

The first LSTM method is a variant of the multi-task healthcare LSTM from Lipton et al. (2016) where the preprocessing involves zero- or last-value carry forward- imputation, mean-reducing, and adding missing indicators. Because our task is nowcasting not multi-label classification, we modify the loss function accordingly. The second LSTM is a WRN preprocessing LSTM, which we compare against to control for differences in WRN preprocessing from that of the first LSTM. The LSTM comprises a linear-embedded input $(i \times h)$ and two LSTM hidden layers $(h \times h)$ that are output to a rectified linear layer $(h \times 1)$ where h is the hidden unit width. For each model, the output is a hazard per time step λ_{ny}, and the loss is the point process log likelihood:

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

The code is written in PyTorch 1.0 and is available at https://github.com/jcweiss2/wrnppl/.

EHR details

From 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 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, with each individual representing a single example. We divided the population into thirds: {train, tune, test} sets. 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. Hemoglobin A1c was measured at least once in 820 individuals (21%), and an adverse event occurred at least once in 137 individuals (3%).

Architecture details

To ease understanding of the relationship between wavelet parameters corresponding to relative time pairwise feature-outcome relationships mapped to absolute time predictions, we provide a diagram corresponding to the wavelet reconstruction networks (WRNs) architecture (Figure A1). The linear form of WRNs facilitates learning a simple and sparse representation that nonetheless captures temporal and value-based associations. Regularization on the transform representation leads to sparse activations which tend to be blockwise and may provide some added robustness to temporal and value-based imprecision.

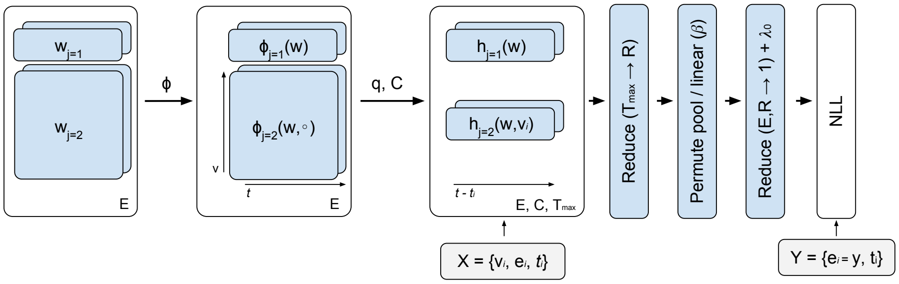

Figure A1: Wavelet reconstruction network architecture. w wavelet parameters, ϕ reconstruction kernel, q relative-to-absolute time translation function, C censoring function, h absolute time functions, β function contribution coefficients, λ_0 baseline rate function.

Additional results

Figure A2 illustrates the ability of WRN to model the rates of complications better than other nowcasting methods, even with a gap in feature availability. Compared to the periodicity seen for A1c measurement, 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.

Multi-forecasting. As one would expect, a trade-off occurs between early prediction and predictive performance. The effect of WRN-PPL forecast distance c on KNR prediction is shown in Figure A2 (right). Notably, the 3-month censored WRN-PPL has approximately the same performance as the nowcasting LSTM[2]. Similarly, effects of single-model, multiple-c prediction are shown in Figure A3 (upper), illustrating the WRN-PPL improvement over WRN for nowcasting (up to $c < 1$) but not for $c \geq 1$. The coefficient profiles vary substantially as a function of c, demonstrating that relative-time attributions, which are commonly used in association statements in health literature, appear to depend on censor time c.

For multi-forecast learning, a comparison of the results in Table 1 and Figure 2 (upper versus lower) demonstrates the value of model expressivity. In particular, Table 1 shows that single forecasting outperforms multi-forecasting at $c = 0$ in Figure A3. However, inspection of Figure 2 (upper versus lower) suggests that multi-forecasting improves the learned wavelet representation insofar as being more self-consistent. Together, these findings suggest that

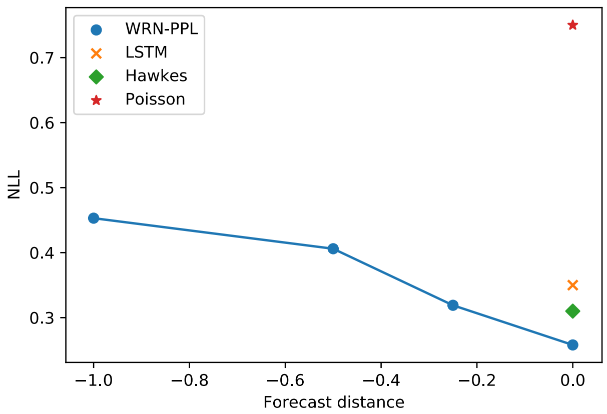

Figure A2: Combined outcome (KNR dataset) negative log likelihood as a function of censor distance (in years).

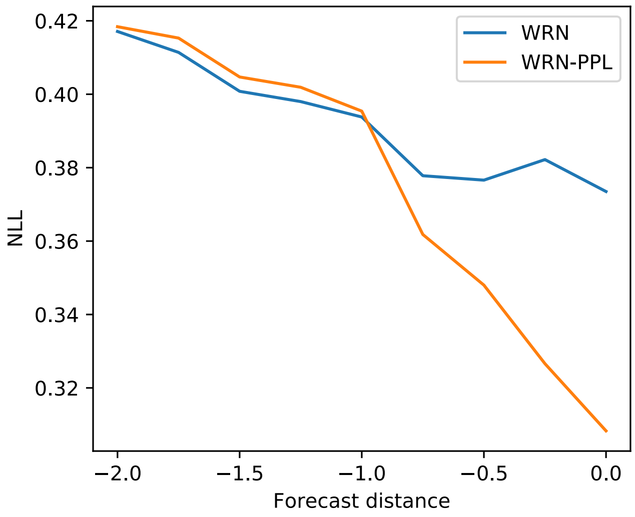

Figure A3: Negative log likelihood (KNR dataset) as a function of forecast censoring distance c for multi-forecasts. The permute and pool layer expresses greater expressivity to model the hazard than WRN.

the layering between the wavelet reconstruction (WRN: {reduction layer, linear}, and WRN-PPL: {reduction, permute and pool, linear}) and the hazard output is not adequately expressive to map the true wavelet reconstruction to the true hazard. We argue the solution is not in simplification nor abandonment of the multi-forecast setting, but in leveraging the multi-forecast setting to facilitate recovery of the wavelet reconstruction by using an

Table A1: Goodman-Kruskal γ, a measure of concordance, on the held out test set. Asterisk (*) denotes simulation.

Dataset	Method (Goodman-Kruskal γ)							
	H. Poisson	Time-invariant	Nowcast	Hawkes	LSTM^1	LSTM^2	WRN	WRN-PPL
ACS*	-1.00	-0.81	0.08	-0.97	**0.98**	0.91	0.80	0.85
A1c*	-1.00	-0.02	0.81	0.64	0.84	**0.85**	0.78	0.77
A1c	-1.00	0.23	0.71	0.79	0.83	0.87	0.93	**0.93**
KNR	-1.00	0.25	0.81	0.91	0.84	0.91	0.95	**0.98**

even more expressive mapping when prediction—rather than associative analysis—is the emphasis.