Learning Dynamic and Personalized Comorbidity Networks from Event Data using Deep Diffusion Processes

Zhaozhi Qian
University of Cambridge

Ahmed M. Alaa
UCLA

Alexis Bellot
University of Cambridge
The Alan Turing Institute

Jem Rashbass
NHS Digital
Public Health England

Mihaela van der Schaar
University of Cambridge
UCLA
The Alan Turing Institute

Abstract

Comorbid diseases co-occur and progress via complex temporal patterns that vary among individuals. In electronic health records we can observe the different diseases a patient has, but can only infer the temporal relationship between each co-morbid condition. Learning such temporal patterns from event data is crucial for understanding disease pathology and predicting prognoses. To this end, we develop *deep diffusion processes* (DDP) to model "dynamic comorbidity networks", i.e., the temporal relationships between comorbid disease onsets expressed through a dynamic *graph*. A DDP comprises events modelled as a multi-dimensional point process, with an intensity function parameterized by the edges of a dynamic weighted graph. The graph structure is modulated by a neural network that maps patient history to edge weights, enabling rich temporal representations for disease trajectories. The DDP parameters decouple into clinically meaningful components, which enables serving the dual purpose of accurate risk prediction and intelligible representation of disease pathology. We illustrate these features in experiments using cancer registry data.

Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108.

1 INTRODUCTION

The illnesses arise not just from individual causes for the specific disease but as a complex interaction between other diseases the patient already had. Identifying and understanding the contribution of comorbidities to disease progression and outcomes is fundamental to medicine and clinical practice. The causal structure of relationships between diseases can be represented by networks that are dynamic in nature. For instance, long-term diabetes increases the risk of cardiovascular and renal disease making high blood pressure and its complications - such as heart attacks more likely. The strength of edges between network nodes changes over time, depending on the entire patient history. Beyond disease progression, these dynamics are also prevalent in economics, finance and sociology (Ahmed & Xing, 2009; Namaki et al., 2011).

In most of these cases, the underlying network dynamics are unknown, and what we observe are sequences of events spreading over the network. To infer the latent network dynamics from observed sequences, one needs to take into account both *when* and *what* events occured in the past since both carry information on the mechanisms involved in disease instantiation and progression. Multi-dimensional point processes are natural candidates for this problem; they explicitly model the time period between events as random variables, and allow them to modulate the *intensity function* — a stochastic model for the time of the next event given previous events. However, traditional parametric models are not expressive enough to capture network dynamics, i.e. the networks they learn are static in nature. On the other hand, existing neural point process models do not entail well-defined network structure due to their complex parameteriza-

tion.

In this paper, we develop the *deep diffusion process* (DDP), a deep probabilistic model for diffusion over comorbidity networks based on mutually-interacting point processes as illustrated in Figure 1. We model the DDP intensity function as an combination of contextualized background risk and networked disease interaction. The disease interaction further consists of three components: (1) static pairwise interactions, (2) time influence, and (3) dynamic influence factors. The first two components are standard in point process models whereas the last component makes use of a deep neural network to (dynamically) update the disease's influence on future events. The introduction of neural networks does not only add to model capacity, but also enables principled predictions based on clinically interpretable parameters which map the patient history on to personalized comorbidity networks. This brings us closer to understanding the underlying disease mechanisms, which as we hypothesize, leads to better out-of-sample and out-of-domain performances. In our experiments, we provide encouraging results in this direction, with better performance of our model in medical data from a different domain.

2 RELATED WORK

In this Section, we highlight previous approaches based on point process formalism, and techniques specifically used in medicine that are relevant to our problem.

2.1 Point Processes for Event Streams

2.1.1 Parametric Models

Gomez-Rodriguez et al. (2012) introduced one of the earliest algorithms for discovering latent networks from sequences of events with a transmission process influenced only by the most recent event. The **cHawkes** model (Choi et al., 2015) removes the Markovian assumption by modelling the event stream as a Hawkes Process where past events temporarily raise the probability of future events. The resulting network captures the pairwise interaction between any two events. However, the influence of the *combinations* of previous events and their *timing* is not accounted for in the network structure. As a result, the learnt network is constant for all event streams at all time. Hence, we refer to it as a *static population-level* network.

2.1.2 Neural Network Based Models

Several recent publications have been focusing on expanding the flexibility of point process models by using recurrent neural networks (RNN).

Du et al. (2016) models the inter-arrival time between consecutive events as a univariate point process and annotates each event with a marker to indicate the event type. Importantly, the marker and the arrival time of the next event are conditionally independent given the history. The independence assumption imposes limitations on the expressiveness of the model as there is only one underlying intensity function for all types of events. **Neural Hawkes** (Mei & Eisner, 2017) models the intensity function directly as a continuous time LSTM. The resulting model has much better flexibility and has achieved the state-of-the-art performance on a variety of prediction tasks. However, the model does not generate a well-defined network between events and it lacks interpretability in general as the hidden state of the RNN do not correspond to clinically meaningful variables.

More recently, the **RPPN** model (Xiao et al., 2019) incorporates temporal attention mechanism to improve the interpretability of neural point process. The model requires a separate attention function for each possible event type in order to connect the observed past with the unobserved future. This may not be an issue when all types of events occur relatively often, but in the medical domain, the majority of diseases have low prevalence in the population[1], which means the attention functions for these diseases may not be adequately trained due to scarcity of data. Lamprier (2019) also considered applying neural networks to information diffusion modelling, although the model does not allow the same type of event to occur more than once. In the medical setting, recurrence of previous diseases carries important information about the patient's health condition. It is also of interest to predict the future recurrence of existing morbidities.

2.2 Medical Disease Networks

Within the medical community, understanding disease networks and associated comorbidities — i.e. any two or more diseases that occur in one person at the same time — is fundamental to the diagnosis and treatment of patients. Many rule-based scoring models are based on empirical association of symptoms and clinical outcomes. For example, the Charlson Comorbidity Index (Charlson et al., 1987) was proposed as early as 1987 to predict the ten-year mortality for a patient by summing up the risk indices associated with various comorbid conditions. The index remains the preferred approach in medical community to represent comorbidity history (Quan et al., 2011).

[1]For example, heart disease is perceived to be very common but it actually occurs in only 1.07% of adults according to official statistics (NHS, 2019a). Rare diseases often have prevalence lower than 0.01%.

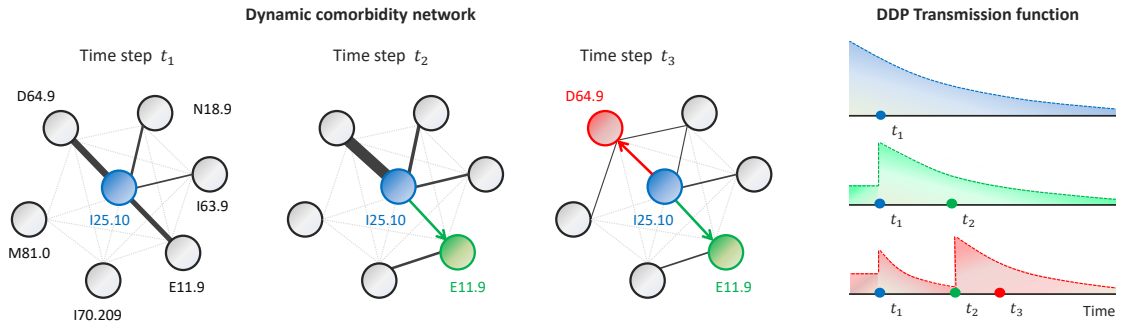

Figure 1: An Exemplary Realization of DDP. Each node corresponds to a disease ICD-10 code (D64.9: Anemia, N18.9: Renal failure, I63.9: Cerebral infarction, E11.9: Diabetes, I70.209: Atherosclerosis, M81.0: Osteoporosis, and I25.10: Heart disease.) Edge weights are depicted via their thickness. The left panels show the evolution of the disease network, and the righmost panel shows the intensity functions of three selected nodes. Onset of heart disease (at t_1) triggers spikes in the intensity functions of diabetes and anemia, making them more likely in the future. The onset of diabetes (at t_2) elevates the risk of anemia (i.e., thicker edge), which consequently occurs at t_3. Edge weights are modulated by a neural network over time.

Table 1: Point Process intensity functions. Subscript v denotes event type. $\boldsymbol{f}$ is the context vector. γ is the time-influence kernel. $\boldsymbol{h}_t$ is the outputs of RNN units and $A_{\cdot,v}$ is the output of event v's attention function. ζ and σ are softplus and sigmoid functions respectively.

Model	Background	Temporal Dependence
Poisson Process	$\mu_v(t)$	0
Hawkes Process	$\mu_v(t)$	$\sum_{i:t_i<t} \alpha_{v_i,v}\gamma_{v_i,v}$
cHawkes	$\mu_v(t,\boldsymbol{f})$	Same as above
Neural Hawkes	0	$\zeta(\boldsymbol{w}_v\boldsymbol{h}_t)$
RPPN	0	$\zeta(\boldsymbol{w}\sum_{i:t_i<t} A_{v_i,v}\boldsymbol{h}_{t_i}\gamma_{v_i,v})$
DDP	$\mu_v(t,\boldsymbol{f})$	$\sum_{i:t_i<t} \alpha_{v_i,v}\gamma_{v_i,v}\sigma(\boldsymbol{w}\boldsymbol{h}_t)$

Recent works have also investigated the construction of data-driven dynamic disease networks (Hu et al., 2019; Lee et al., 2019; Beck et al., 2016; Hidalgo et al., 2009). However, with no exception, the networks in these works are constructed in two steps. First, certain pairs of diseases are linked together based on population level statistics such as risk ratio or temporal correlation. Next, the disease pairs are pieced together into longer trajectories or networks. Since all the information used in this process is on population level, the resulting graph is not *personalized*. Furthermore, constructing the network by combining pieces usually implies strong independence assumptions e.g. Markovian assumption, which rarely holds in disease progression (a real example is given in the appendix). Therefore, the *dynamic* aspect of disease progression is not adequately represented.

The main contribution of this work is to augment the above approaches by modelling the disease network itself as an individualized dyanmic graph. This allows us to model more complex temporal interactions between diseases as well as provide personalized predictions.

3 DEEP DIFFUSION PROCESS

3.1 Dynamic Network Representation

Consider a dynamic network $G_t = (V, L_t, E_t, W_t)$ consisting of a set of vertices $V = \{1, ..., K\}$ annotated with binary labels $L_t = \{0,1\}^K$, and a set of directed edges E_t weighed by $W_t = \mathbb{R}^{+|E|}$. The vertices V correspond to the set of all possible event types. At any time, a vertex v has label $l_v = 1$ if a type v event has occurred or $l_v = 0$ otherwise. The edge set, formally defined as $E \subseteq \{(v_i, v_j)|v_i, v_j \in V, l_{v_i} = 1\}$, contains edges that link an observed vertex v_i to another vertex v_j if v_i modulates v_j's chance of occurrence. The edge weights W_t represent the strength of such modulation effect between events (Refer to Figure 1).

While we do not observe the network directly at each time, we do have access to individual trajectories through the network, available as a sequence of events and corresponding time points,

$$\mathcal{H} = \{(t_1, v_1), ..., (t_n, v_n)\} \tag{1}$$

where $t_i \in \mathbb{R}^+$ is the time of occurrence and $v_i \in V$ is the associated event type. From the event sequence, one can immediately derive the vertex label L_t for $t \leq t_n$. Since the vertex set V is fixed a-priori depending on the problem scope, the remaining unknown components of the graph are the label L_t for $t > t_n$ as well as the weighted edges E_t, W_t for $t > 0$.

Determining the future vertex label is known as *event prediction*, whereas uncovering weighted edges corresponds to *network inference*. Our goal is to devise a model that addresses both problems simultaneously.

3.2 Preliminaries on Point Process

Before formally introducing the DDP model, we first recapitulate several key concepts of point process.

Lying at the core of point processes is the *intensity function* $\lambda_v(t)$, which is the probability of event v occurring in time window $[t, t+dt)$ given a history $\mathcal{H}_t := \{(t_i, v_i) : t_i < t\}$, *i.e.*,

$$\lambda_v(t)dt := Pr(\text{event of type } v \text{ in } [t, t+dt)|\mathcal{H}_t) \quad (2)$$

As we can see in Table 1, different point process models have different parameterizations of the intensity function ranging from the simplest Poisson process to the complex Neural Hawkes and RPPN. However, once the intensity function is given, many interesting properties can be readily derived. For example, the likelihood of an observed sequence $\mathcal{H}_T$ is given by

$$\mathcal{L}(\theta; \mathcal{H}_T) := \prod_{(t_i, v_i) \in \mathcal{H}_T} \lambda_{v_i}(t_i) \exp\left(-\int_{t_{i-1}}^{t_i} \lambda(\tau)d\tau\right), \quad (3)$$

where $\lambda(t) = \sum_{v \in \mathcal{V}} \lambda_v(t)$ and θ is the collection of all free parameters in the model. As another example, the probability of an type v event happening at a *specific* time $t_{i+1} > t$ is given by

$$P(v_{i+1} = v \,|\, \mathcal{H}_t, t_{i+1}) = \frac{\lambda_v(t_{i+1})}{\lambda(t)} \quad (4)$$

and the occurrence time of the next event is given by

$$P(t_{i+1} = t \,|\, \mathcal{H}_T) = \lambda(t) \exp\left(-\int_{t_i}^{t} \lambda(\tau)d\tau\right) \quad (5)$$

The model can be trained in multiple ways. In general, one can maximize the likelihood function 3 via stochastic gradient descent. If the integral term does not have a closed form, it can be approximated by Monte Carlo sampling as done in Mei & Eisner (2017). In addition, it is also possible to train the model by minimizing the prediction loss based on (4) and (5) (Du et al., 2016).

3.3 Model Specification

This section presents the Deep Diffusion Process — a deep probabilistic model for inferring network dynamics while accurately predicting future events.

3.3.1 intensity function

Our objective is to enrich the intensity function in order to capture the time-dependent disease-to-disease relationships. To this end, we decompose the overall intensity function into two additive components:

$$\lambda_v(t) = \mu_v(\boldsymbol{f}) + \sum_{t_i < t} g_v(v_i, t_i, \mathcal{H}_{t_i}, t). \quad (6)$$

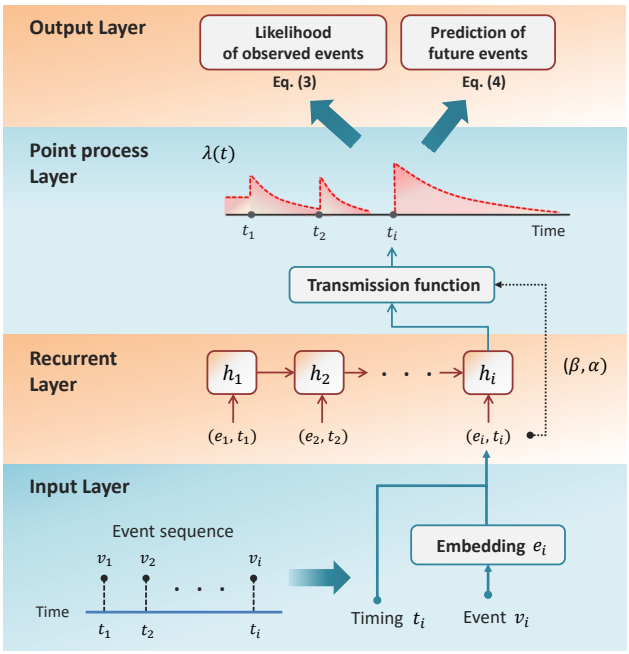

Figure 2: Schematic of the DDP Architecture.

The first term captures the occurrence of events due to static exogenous risk factors $\boldsymbol{f}$. For example, it can model the increased risk of heart attack among the obese patients. The second term models the impact from past events. Each historical event adds an "impulse" g_v to the intensity function of event v depending on the event type v_i, the timing t_i, and, most importantly, the event history $\mathcal{H}_{t_i}$ at the time. This decomposition allows modeling the impact of exogenous factors and that of past events separately.

Next, we introduce the parametric form to capture the impact from past events as follows:

$$g_v(v_i, t_i, \mathcal{H}_{t_i}, t) = \alpha_{v_i,v} \cdot \gamma_{v_i,v}(t - t_i) \cdot w_i(\mathcal{H}_{t_i}). \quad (7)$$

The parameter $\alpha_{v_i,v}$ captures the instantaneous impact from event v_i to v. We use α to denote the matrix that contains all the $\alpha_{v_i,v}$ between any two events. The time influence kernel $\gamma_{v_i,v}(t - t_i)$ captures the decay or increase of previous events' influence. It is a non-negative function defined on $\mathbb{R}^+$ and integrates to one. One common choice is the exponential kernel $\gamma_{v_i,v}(t - t_i) = \beta_{v_i,v} \exp(-\beta_{v_i,v}(t - t_i))$.

The last component, $w_i(\mathcal{H}_{t_i}) \in [0, 1]$, is the (dynamic) influence factor that depends on the full patient history. It is learned by an RNN applied to the sequence of past events. As shown in Figure 2, the event v_i is encoded through an embedding $e_i \in \mathbb{R}^D$, which together with the time gap $\Delta t_i := t_i - t_{i-1}$ are fed into the recurrent layer as follows:

$$h_i = LSTM(e_i \odot \Delta t_i, h_{t-1}), \quad (8)$$

where h_i is the LSTM output. In our implementation, we used standard LSTM with the time gap as an additional input dimension as shown in (8). However, we note that any continuous-time RNN (e.g., phased LSTM (Neil et al., 2016)) is applicable. The influence factor w_i is then given by

$$w_i = \sigma(h_i \cdot W + b), \tag{9}$$

where W and b are parameters to be learned, and $\sigma(.)$ is the sigmoid function.

For training, we use a loss function comprising the likelihood function $L(\mathcal{H}_T;\theta)$ in (3) and the cross entropy loss $l_p(\mathcal{H}_T;\theta)$ for event type prediction, i.e.,

$$\theta^* = \text{argmax}\, \mathcal{L}(\mathcal{H}_T;\theta) - \eta \cdot l_p(\mathcal{H}_T;\theta) \tag{10}$$

where $\eta > 0$ is a hyperparameter that trade -off the two objectives, and is determined from a validation set. The loss function in (10) encapsulates our dual objective of a faithful representation for the observed event sequence and the ability to predict the next event.

3.4 Dynamic Network Inference

The parameter α, the time influence kernel γ and the influence factors w_i jointly define the network structure at time t. α is the baseline matrix that encodes static pairwise relationships. The larger the value of α_{uv}, the more influence event u will have on event v *on average*. The time influence kernel further modulates the link strength based on the time gap between the occurrence of events.

The influence factor w_i modulates α and enables the network structure to adapt to the observed event sequence. Based on the full history of past events, the influence factor may strengthen or diminish the impact of one particular event and thus modifies all its outgoing links. Therefore, at time t, the directed edge $v \to u$ will have weight

$$W_{v\to u}(t) = \sum_{\substack{(t_i,v_i)\in\mathcal{H}_T \\ v_i=v}} \alpha_{v,u} \cdot \gamma_{v,u}(t-t_i) \cdot w_i. \tag{11}$$

It is worth highlighting that the resulting graph is dynamic in two aspects. First, the influence factor w_i is updated for each event v_i based on the full event history up to that point. Depending on the combination and the timing of historical events, the influence factor for subsequent events will differ, leading to a different graph structure. Secondly, the time influence kernel γ modulates the edge weight as time moves on.

It is often desirable for the graph to have a sparse structure i.e. $W_{v\to u} = 0$ for many pairs of events v, u. We can introduce a L_1 regularization term for α matrix to encourage sparsity as proposed in Choi et al. (2015).

Lastly, we note that sometimes it is required to construct a population-level static graph instead of a dynamic graph to capture the high-level event interaction. Static edge weights can be found by averaging out the dynamic components in equation 11 as follows

$$W_{v\to u} = \alpha_{v,u}\mathbb{E}[w_i], \tag{12}$$

where the expectation represents the average influence factor of event v.

4 EXPERIMENTS

In this Section, we utilize data from a large-scale cancer registry to evaluate DDP[2]. Throughout our experiments, we evaluate DDP with respect to three aspects: (a) its ability to extract interpretable disease networks that are sensible in the light of current medical literature (Section 4.2), (b) its accuracy in predicting disease pathways (Section 4.3), and (c) its generalizability to out-of-domain datasets (Section 4.4).

4.1 Data Description

We used national registry data for a cohort of colorectal cancer patients diagnosed between 2011 and 2015. The data comprises 268,000 observations of 100 common diagnoses for 54,000 patients. Each patient is associated with up to 15 comorbidities. The earliest diagnoses date back to the 1990s, which gives us a fairly broad timescale to study the progression of colorectal cancer. In addition to the primary dataset described above, we have also considered data for 25,000 patients with stomach cancer. It is well understood that patients with stomach cancer are exposed to different risk factors from patients with colorectal cancer (Miller, 1982; Drasar & Irving, 1973). Therefore, we use this dataset as an **out-of-domain** test set to validate transferability and robustness of DDP.

4.2 Colorectal cancer comorbidity networks

Heterogeneity of disease pathways among patients can be quantified by measuring the distance between their comorbidity networks. A commonly used metric for measuring distance between graphs is the Jaccard index as illustrated in Figure 3 (a) (Real & Vargas, 1996). To handle weighted graphs, we use the weighted Jaccard index $J(X,Y) = \frac{\sum_i \min(x_i,y_i)}{\sum_i \max(x_i,y_i)}$ where x_i and y_i are the edge weights in two graphs X and Y. Within a population, the average Jaccard distance J_{avg} between any two individual networks measures the *heterogene-*

[2]Implementation details are provided in the appendix

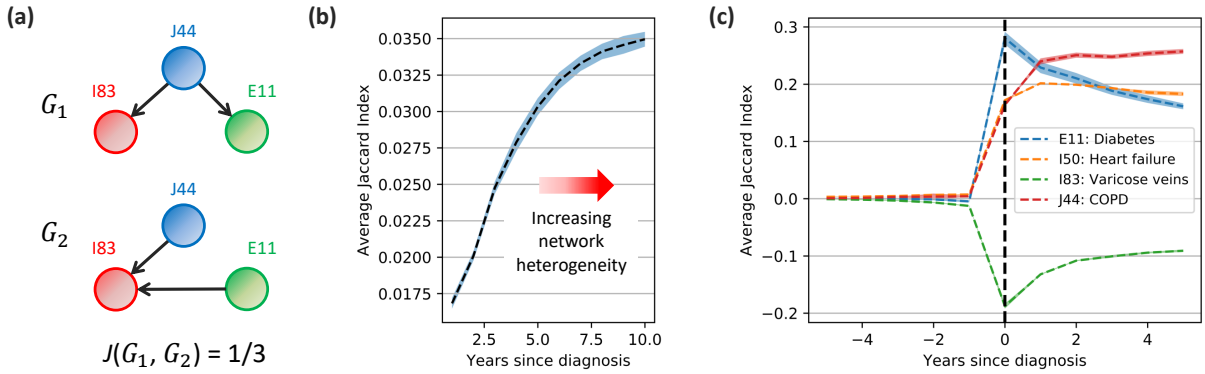

Figure 3: (a) Illustration of the Jaccard index between two exemplary graphs. (b) heterogeneity of comorbidity networks increases over time in colorectal cancer pathways. (c) Onset of a comorbidity modulates future pathways.

ity of the population, i.e.,

$$J_{avg} = \frac{\sum_{n<m} 1 - J(G_n, G_m)}{\binom{N}{2}}, \tag{13}$$

where N denotes the size of population. The larger the average distance, the more spread out the population. Since the disease networks are time-varying, we compute $J_{avg}(t)$ based on the the networks at time t to reflect the heterogeneity at that moment.

Patient pathways get more heterogeneous over time. In Figure 3 (b), we track J_{avg} over time (referenced to the date of initial diagnosis). We can readily see that the comorbidity networks learned by DDP become increasingly heterogeneous as time progresses. This reflects the fact that as a patient gets older, more comorbidities will occur and the subsequent disease pathway will become more complex. The increase in heterogeneity also highlights the need for modeling comorbidities with a personalized method since any one-size-fits-all approach will under-appreciate the diversity occurring later in the pathway. The ability of DDP to accurately predict the heterogeneity of colorectal cancer pathways is assessed in Section 4.3.

Figure 3 (c) shows how the network dynamically adapts to "influencers", diseases which trigger a large variety of comorbidities and complications. The figure displays change in J_{avg} before and after a disease onset relative to the population average. Positive value means increase in heterogeneity relative to the population, negative value otherwise, and zero means no change. Time is normalized so that the disease of interest always occurs at time 0. The red, blue and orange lines represent chronic obstructive pulmonary disease (COPD), Type 2 diabetes mellitus and heart failure respectively. It is well-established in the medical literature that all three types of diseases have complex heterogeneous comorbidity pathway (Fabbri et al., 2008; Stratton et al., 2000). This is clearly reflected in Figure 3 (c) where the onset of these diseases triggers an immediate and persistent increase in the heterogeneity of comorbidity networks. On the other hand, the green line represents varicose veins of lower extremities, a mild condition that often does not need treatment (NHS, 2019b). It is thus unsurprising to see that patients with this condition usually have less heterogeneous disease networks than the average.

The above analysis shows DDP's ability to adjust the subsequent comorbidity pathway based on the occurrence of individual diseases. This high resolution view can help medical researchers better understand the progression and taxonomy of diseases.

Individual-level comorbidity networks. Figure 4 depicts the evolution of the dynamic comorbidity network of five common gastrointestinal disorders (for one patient's pathway) as inferred by DDP — the comorbidities included: diverticular diseases (ICD-10 code K57), intestinal disorders (K63), benign neoplasm in the colon and rectum (D12), diverticular diseases (K57), and ulcerative colitis (K52). The intensity function corresponding to the patient's trajectory is shown in the bottom panel. Each edge's thickness in the network corresponds to the likelihood of the disease designating the receiving node to occur at a given time step. The individual patient's dynamic network is contrasted with a static network (upper right panel) constructed directly from raw data by counting the co-occurrences of each pair of comorbidities and weighting edges accordingly.

As we can see in Figure 4, the DDP comorbidity network is fairly dense at each time step, which suggests that the diseases are related. In fact, numerous medical publications have examined associations between these diseases. For example, it has been established that a lack of dietary fiber intake underlies the onset of

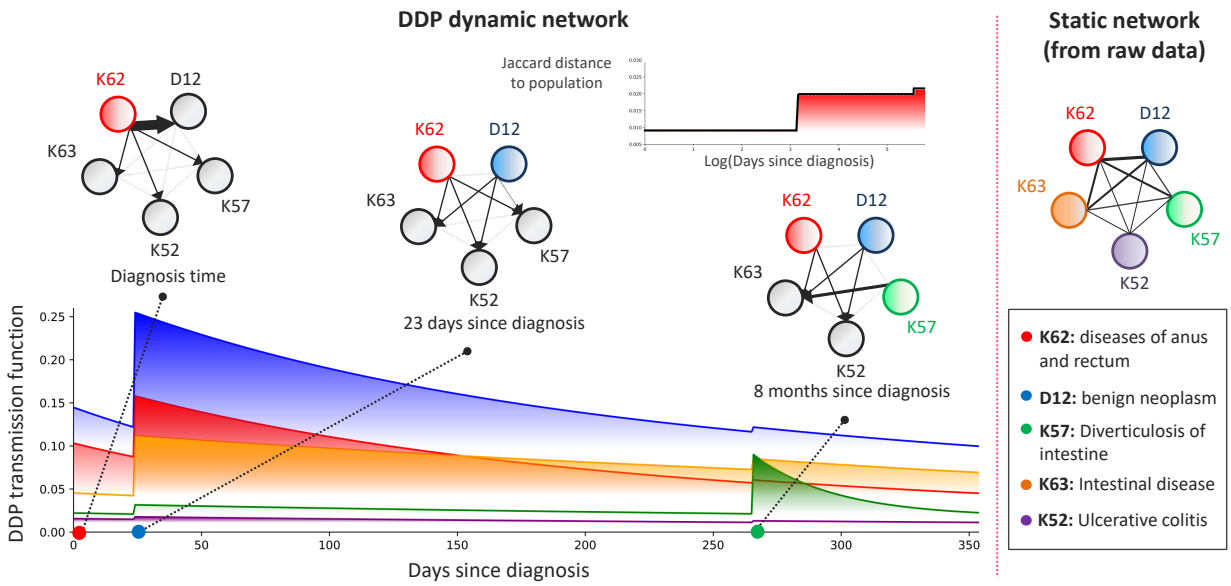

Figure 4: The dynamic comorbidity network learned by DDP for an individual patient at three time steps, together with the corresponding intensity function. Nodes for diseases that have not occurred are colored in gray, and disease already diagnosed are assigned a distinct color. Edge thickness correspond to the disease likelihood at the given time step. In the upper left panel, we plot the Jaccard distance of the patient's network with respect to the average population as a function of time (on a logarithmic scale). The static comorbidity network obtained by counting disease co-occurrences and using the counts as graph edges is depicted on the right panel.

diverticular diseases, intestinal disorders, and tumours of the colon and rectum (Painter & Burkitt, 1971; Burkitt et al., 1972). Furthermore, there are strong epidemiological evidences of associations between tumours of colon and rectum, diverticular diseases, and ulcerative colitis (Ekbom et al., 1990; Burkitt, 1971).

Figure 4 shows that the dynamics of the inferred individualized comorbidity network cannot be deduced from the approach presented in Beck et al. (2016). That is, at each time step the RNN component of the DDP adapts the weights on the network edges to reflect the impact of previous diagnoses on the odds of future ones. Moreover, the weights of the network edges reflect the *timing* at which future comorbidities are expected to occur — for instance, D12 occurs only 23 days after diagnosis, which was correctly anticipated by the DDP network as it assigned a large weight to the edge connecting the (pre-existing comorbidity) K62 and D12 at diagnosis time. On the contrary, K57 occured more than 8 months after diagnosis, which also was anticipated by DDP model having assigned a smaller weight to edges flowing into K57 node.

By measuring the average (Jaccard) distance between the comorbidity network of the patient at hand and those of the overall patient population (upper left panel of Figure 4), we can see that the patient's network diverges from the typical population-level pathway as time progresses. This emphasizes the importance of the personalization aspect of the DDP model in predicting patient prognosis in later stages of the disease as we will show in the next Section.

4.3 Predicting colorectal cancer pathways

Prediction targets and evaluation metric. Each individual patient has a unique disease pathway and a good representation of their health trajectory should enable differentiating these pathways. We evaluate how well DDP discerns the future disease pathways by predicting the next event. Given a disease history $\mathcal{H}_t = \{(t_i, v_i) : t_i < t\}$, the models try to predict the probability of having a disease v_{i+1} at time t_{i+1}. The time t_{i+1} represents the time of the next disease onset available in the dataset. We chose to predict the incidental risk at a given time due to the nature of our data. For most chronic diseases including cancer, the diagnosis may occur much later than the actual onset. Hence the true disease onset time as well as the time between disease onsets are never observed. By focusing on predicting the diseases at known diagnosis time, evaluation becomes more objective and less prone to unknown variation. We calculate the Area Under ROC (AUC) score for predicting prevalent comorbidities.

Benchmarks. We compare the performance of DDP with *Neural Hawkes* (Mei & Eisner, 2017), *cHawkes* (Choi et al., 2015), *Charlson Score* (Charlson et al., 1987), and RETAIN (Choi et al., 2016), which is a recurrent neural network with temporal attention mech-

Table 2: AUC (± 95% confidence intervals) performance for all baselines. Best performance is highlighted in bold font.

ICD-10 code	DDP	Neural Hawkes	cHawkes	Charlson	RETAIN
I50	**0.74 ± 0.0114**	0.72 ± 0.0127	0.69 ± 0.0136	0.68 ± 0.0111	**0.73 ± 0.0123**
N39	**0.64 ± 0.0085**	0.62 ± 0.0085	0.59 ± 0.0083	0.58 ± 0.0079	**0.65 ± 0.0085**
A41	**0.72 ± 0.0091**	0.72 ± 0.0092	0.71 ± 0.0091	0.60 ± 0.0098	0.70 ± 0.0101
D12	**0.69 ± 0.0053**	0.67 ± 0.0055	0.66 ± 0.0055	0.59 ± 0.0055	0.66 ± 0.0059
E86	**0.72 ± 0.0225**	**0.72 ± 0.0235**	0.69 ± 0.0222	0.52 ± 0.0222	0.58 ± 0.0222
I25	**0.79 ± 0.0081**	0.77 ± 0.0089	0.77 ± 0.0087	0.63 ± 0.0085	0.77 ± 0.0084
K63	**0.68 ± 0.0061**	0.64 ± 0.0064	0.64 ± 0.0063	0.60 ± 0.0060	0.65 ± 0.0065
K83	**0.69 ± 0.0217**	0.68 ± 0.0225	0.66 ± 0.0209	0.63 ± 0.0200	0.62 ± 0.0224

Table 3: Out-of-domain AUC performance for all baselines. Best performance is highlighted in bold font.

ICD-10 code	DDP	Neural Hawkes	cHawkes	Charlson	RETAIN
I50	**0.73 ± 0.0089**	0.69 ± 0.0091	0.65 ± 0.0102	0.67 ± 0.0195	0.71 ± 0.0093
N39	**0.65 ± 0.0063**	0.56 ± 0.0066	0.59 ± 0.0064	0.62 ± 0.0142	0.63 ± 0.0066
A41	**0.69 ± 0.0065**	0.59 ± 0.0070	0.65 ± 0.0070	0.62 ± 0.0142	0.66 ± 0.0072
D12	**0.68 ± 0.0065**	0.56 ± 0.0068	0.66 ± 0.0065	0.57 ± 0.0147	0.63 ± 0.0078
E86	**0.65 ± 0.0127**	0.52 ± 0.0125	0.62 ± 0.0121	0.55 ± 0.0321	0.56 ± 0.0122
I25	**0.78 ± 0.0049**	0.66 ± 0.0058	0.75 ± 0.0054	0.59 ± 0.0117	0.75 ± 0.0053
K63	**0.65 ± 0.0077**	0.58 ± 0.0079	0.60 ± 0.0078	0.57 ± 0.0164	0.63 ± 0.0083
K83	**0.69 ± 0.0126**	0.60 ± 0.0135	0.65 ± 0.0123	0.57 ± 0.0284	0.63 ± 0.0128

anism akin to RPPN (Xiao et al., 2019). Section 2 contains a detailed review of these models.

Results. The results are shown in Table 2. In five out of eight cases DDP achieved the best performance. In the rest three cases, the performance of DDP is comparable to that of Neural Hawkes or RETAIN. However, in these three cases, DDP does not only offer a competitive predictive accuracy, but infers the comorbidity network as well — comorbidity networks cannot be straightforwardly inferred from the parameters of Neural Hawkes and RETAIN. This does not only provide more elaborate interpretability, but as we show in Section 4.4, it enables better generalization to out-of-domain data as we will. In addition, we can readily see that DDP always outperforms cHawkes and the Charlson Score. This suggests that the history-independent triggering mechanics of cHawkes and Charlson score do not adequately capture the disease complexity.

4.4 Transferability to other types of cancer

Finally, we applied all baselines originally trained on the primary dataset (colorectal cancer) to the out-of-domain dataset (stomach cancer) *without* re-training. All other aspects of the experimental setup remains the same as the previous Section. The results are illustrated in Table 3. We can clearly see that DDP outperforms all the benchmarks including Neural Hawkes by a big margin on the out-of-domain samples.

We performed a post-hoc analysis to better understand what the DDP model has learned. First, we performed Chi-squared tests to test whether the prevalence of individual diseases or the occurrence of disease pairs are different in the two data sets. In both cases, the test concluded that the distributions are different.(p-value < 0.001). This finding suggests that the disease networks constructed based on population level statistics such as those reviewed in Section 2.2 will tend to be **different** for the two datasets. Next, we randomly sampled a subset of patients from each of the two datasets and calculated Jaccard distance within and between the datasets. Student's t-test concluded the average distance between two groups are smaller than the distance within the group (p-value < 0.01). This indicates that the graph heterogeneity across datasets are smaller than the one within. In other words, the disease network learned by DDP applies to both sets of patients and is generalizable across cancer sites.

5 CONCLUSION

In this paper, we developed DDP that utilizes deep neural networks to enable both *accurate* prediction of disease trajectories and *interpretable* representations of disease pathways. Our experiments show that DDP can offer more nuanced understanding of disease progression mechanisms, more accurate prediction of patient pathways, and better generalizability across different diseases. By taking into account the full disease history, the learned DDP comorbidity networks are well equipped to deal with individual-level disease trajectories in a data-driven fashion, improving over existing one-size-fits-all clinical guidelines.

ACKNOWLEDGEMENTS

The support and advice of the analytical and registration staff in Public Health Englands National Cancer Registration Service. Only completely anonymous data was used in this work.

References

Amr Ahmed and Eric P Xing. Recovering time-varying networks of dependencies in social and biological studies. *Proceedings of the National Academy of Sciences*, 106(29):11878–11883, 2009.

Mette K Beck, Anders Boeck Jensen, Annelaura Bach Nielsen, Anders Perner, Pope L Moseley, and Søren Brunak. Diagnosis trajectories of prior multimorbidity predict sepsis mortality. *Scientific reports*, 6:36624, 2016.

Denis P Burkitt. Epidemiology of cancer of the colon and rectum. *Cancer*, 28(1):3–13, 1971.

Denis P Burkitt, ARP Walker, and No S Painter. Effect of dietary fibre on stools and transit-times, and its role in the causation of disease. *The Lancet*, 300 (7792):1408–1411, 1972.

Mary E Charlson, Peter Pompei, Kathy L Ales, and C Ronald MacKenzie. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *Journal of chronic diseases*, 40(5):373–383, 1987.

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

Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. In *Advances in Neural Information Processing Systems*, pp. 3504–3512, 2016.

BS Drasar and Doreen Irving. Environmental factors and cancer of the colon and breast. *British Journal of Cancer*, 27(2):167, 1973.

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*, pp. 1555–1564. ACM, 2016.

Anders Ekbom, Charles Helmick, Matthew Zack, and Hans-Olov Adami. Ulcerative colitis and colorectal cancer: a population-based study. *New England journal of medicine*, 323(18):1228–1233, 1990.

LM Fabbri, Fabrizio Luppi, Bianca Beghé, and KF Rabe. Complex chronic comorbidities of copd. *European Respiratory Journal*, 31(1):204–212, 2008.

Manuel Gomez-Rodriguez, Jure Leskovec, and Andreas Krause. Inferring networks of diffusion and influence. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 5(4):21, 2012.

Csar A. Hidalgo, Nicholas Blumm, Albert-Lszl Barabsi, and Nicholas A. Christakis. A dynamic network approach for the study of human phenotypes. *PLoS computational biology*, 5(4):e1000353, April 2009. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1000353.

Jessica X Hu, Marie Helleberg, Anders B Jensen, Søren Brunak, and Jens Lundgren. A large-cohort, longitudinal study determines precancer disease routes across different cancer types. *Cancer research*, 79(4):864–872, 2019.

Sylvain Lamprier. A recurrent neural cascade-based model for continuous-time diffusion. In *International Conference on Machine Learning*, pp. 3632–3641, 2019.

Dong-Gi Lee, Myungjun Kim, and Hyunjung Shin. Inference on chains of disease progression based on disease networks. *PloS One*, 14(6):e0218871, 2019. ISSN 1932-6203. doi: 10.1371/journal.pone.0218871.

Hongyuan Mei and Jason M Eisner. The neural hawkes process: A neurally self-modulating multivariate point process. In *Advances in Neural Information Processing Systems*, pp. 6754–6764, 2017.

AB Miller. Risk factors from geographic epidemiology for gastrointestinal cancer. *Cancer*, 50(11 Suppl): 2533–2540, 1982.

A Namaki, AH Shirazi, R Raei, and GR Jafari. Network analysis of a financial market based on genuine correlation and threshold method. *Physica A: Statistical Mechanics and its Applications*, 390(21-22): 3835–3841, 2011.

Daniel Neil, Michael Pfeiffer, and Shih-Chii Liu. Phased lstm: Accelerating recurrent network training for long or event-based sequences. In *Advances in neural information processing systems*, pp. 3882–3890, 2016.

NHS. Condition prevalence. `https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/general-practice-data-hub/condition-prevalence`, 2019a. Accessed: 2019-09-30.

NHS. Treatment for varicose veins. `https://www.nhs.uk/conditions/varicose-veins/treatment/`, 2019b. Accessed: 2019-09-30.

Neil S Painter and Denis P Burkitt. Diverticular disease of the colon: a deficiency disease of western civilization. *British medical journal*, 2(5759):450, 1971.

Hude Quan, Bing Li, Chantal M Couris, Kiyohide Fushimi, Patrick Graham, Phil Hider, Jean-Marie Januel, and Vijaya Sundararajan. Updating and validating the charlson comorbidity index and score for risk adjustment in hospital discharge abstracts using data from 6 countries. *American journal of epidemiology*, 173(6):676–682, 2011.

Raimundo Real and Juan M Vargas. The probabilistic basis of jaccard's index of similarity. *Systematic biology*, 45(3):380–385, 1996.

Irene M Stratton, Amanda I Adler, H Andrew W Neil, David R Matthews, Susan E Manley, Carole A Cull, David Hadden, Robert C Turner, and Rury R Holman. Association of glycaemia with macrovascular and microvascular complications of type 2 diabetes (ukpds 35): prospective observational study. *Bmj*, 321(7258):405–412, 2000.

Shuai Xiao, Junchi Yan, Mehrdad Farajtabar, Le Song, Xiaokang Yang, and Hongyuan Zha. Learning time series associated event sequences with recurrent point process networks. *IEEE transactions on neural networks and learning systems*, 2019.