Beyond Topics: Discovering Latent Healthcare Objectives from Event Sequences

Adrian Caruana*, Madhushi Bandara, Daniel Catchpoole, Paul J Kennedy

Australian Artificial Intelligence Institute, Faculty of Engineering and IT, University of Technology Sydney, Sydney Australia

INTRODUCTION

A strong understanding of how patients are treated in a healthcare setting is fundamental to improving patient outcomes. However, the objectives, intentions, and decision-making of healthcare practitioners is often loosely defined and not recorded in electronic health records (EHR). The theoretical goal of this paper is to develop a method for learning about high-level concepts, such as best-practice guidelines, clinical research, and centralised healthcare management, from low-level data, such as standardised disease and drug codes.

To facilitate this goal, we introduce an abstraction layer, referred to as **healthcare objective**. Healthcare objectives encapsulate reasoning behind the formation of particular EHR sequences. The technical goals of this paper are to define the characteristics of healthcare objectives, simulate them and subsequently rediscover them in data, and apply the methods to real-world datasets. We justify why our novel method, **Ca**tegorical **S**equence **E**ncoder (CaSE), is necessary to discover latent healthcare objectives, and demonstrate the characteristics that CaSE captures from real-world EHRs.

CONTRIBUTIONS OF THIS PAPER

- Characterisation of healthcare objectives, and prerequisites for identifying them from EHR sequences.
- Description of a synthetic data model to model healthcare objectives.
- Introduction of Categorical Sequence Encoding (CaSE), a generalised methodology for generating representations of categorical sequences.
- Experimental validation of healthcare objective identification in synthetic and authentic EHR data.

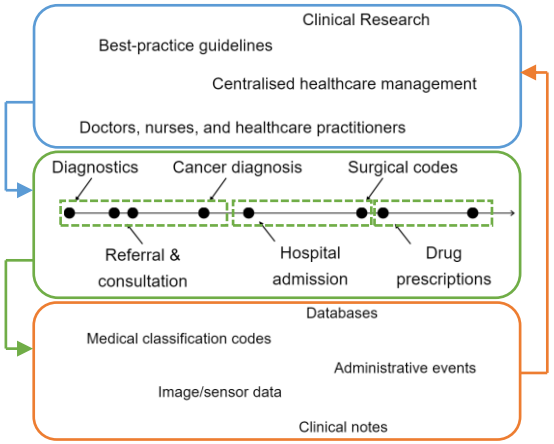

Fig. 1: From bottom-to-top, this picture depicts increasingly abstract healthcare information. By using low-level EHR data (orange), this research seeks to learn the high-level clinical concepts (blue) that affect how patient treatment progresses in a healthcare environment. The dashed-rectangles (green) indicate an intermediate abstraction, healthcare objectives, to facilitate learning about the high-level concepts.

* Corresponding author: adrian.caruana@uts.edu.au

METHODOLOGY

MODELLING LATENT TREATMENT GROUPS IN ELECTRONIC HEALTHCARE RECORDS

Treatment events are categorical samples from the discrete set of all possible treatment events $\mathbb{E}$. Let x be a sample in an EHR dataset where $X \in \mathbb{E}$ such that $X \sim P$ with P a discrete probability distribution over the set $\mathbb{E}$. In practice, P is not uniformly distributed and depends on the healthcare objective being applied. Each healthcare objective will alter the distribution of observed treatment events, resulting in a **treatment group** g. The distribution of treatment events is given by $P(X, G)$, and each event is sampled based on which treatment group g is being expressed from a set of possible treatment groups G where $g \in G$. We seek to construct a representation $\hat{P}$ that approximates P. The goal of this methodology is to identify areas of high local density in $\hat{P}$ to infer the existence latent treatment groups $G \in \mathbb{G}$.

SYNTHETIC ELECTRONIC HEALTHCARE RECORDS

We implement a synthetic data model defining a set of possible treatments $\mathbb{E}$, a set of treatment groups $\mathbb{G}$, and yields observations x drawn from a discrete probability distribution $P(X, G)$ where $X \in \mathbb{E}$ and $G \in \mathbb{G}$. The distribution of P for a particular treatment group g is $P(X \mid G = g) \sim \text{Zipf}(\beta_g)$. Additionally, each g corresponds to a random permutation over the set $\mathbb{E}$ as shown in **Fig. 2**.

Using these definitions, synthetic patient treatment sequences **(1)** are generated with latent treatment groups **(2)**. Each treatment group is sampled according to **(3)**, and each treatment event is sampled according to **(4)**. Repeating this process many times yields an EHR dataset of synthetic patient treatment sequences, as depicted in **Fig. 3**.

$$x_1, x_2, x_3, \ldots x_i, \ldots x_n \quad (1)$$

$$g_1, g_2, g_3, \ldots g_i, \ldots g_n \quad (2)$$

$$P(G = g_i | Q = q) = \begin{cases} g \leftarrow P(G) & q < \alpha \\ g_{i-1} & q \geq \alpha \end{cases} \quad (3)$$

$$x_i \leftarrow P(X | G = g_i) \quad (4)$$

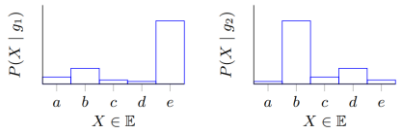

Fig. 2: The distribution of treatment events $X \in \{a, b, c, d, e\}$ given a latent treatment group g_i , with $|\mathbb{G}| = 2$ and $|\mathbb{E}| = 5$. Each group $G \in \mathbb{G}$ randomly permutes $\mathbb{E}$, with the distribution being $Zipf$.

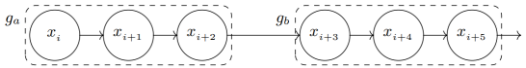

Fig. 3: The diagram depicts a sequence of observed treatment events x (circles) and the latent treatment groups g (rectangles).

LEARNED TREATMENT GROUP REPRESENTATIONS

We learn representations of latent treatment groups in EHRs using a two-step encoding process called Categorical Sequence Encoding (CaSE). In the first stage, **Cat2Vec**, an encoding representation of healthcare events is learned using a Siamese Network as in **Fig. 4a**. In the second stage, **Seq2Seq**, an encoding representation of patient treatment sequences is learned using a Transformer Network as in **Fig. 4b**.

EXPERIMENTS

EMPIRICAL DEMONSTRATION

To demonstrate how CaSE learns latent treatment groups in EHRs, a synthetic dataset with $|\mathbb{G}| = 6$ and $|\mathbb{E}| = 100$ is used. The learned encoding representations from Cat2Vec and Seq2Seq are visualised using a UMAP embedding in **Fig. 5a** and **Fig. 5b** respectively. As **Fig. 5** depicts, **Cat2Vec** is not sufficient to identify which treatment events belong to the same treatment group, whereas **Seq2Seq** groups distinct treatment events together in the learned representation.

SYNTHETIC DATA

We evaluate CaSE for classifying latent treatment groups in synthetic data. We vary $|\mathbb{G}|$ and $|\mathbb{E}|$ in the synthetic data configuration and use with Latent Dirichlet Allocation (LDA) as a baseline. A post-hoc clustering (PHC) is used on CaSE encodings to classify treatment groups. CaSE is shown to outperform LDA at event-level classification of treatment groups, regardless of $|\mathbb{G}|$ and $|\mathbb{E}|$, as shown in **Table 1**.

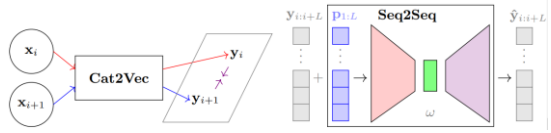

Fig. 4a Fig. 4b

Fig. 4: Fig. 4a depicts the Siamese Network, which learns to minimises the distance (violet) between adjacent events (red and blue) encoded to the latent vector space. **Fig. 4b** depicts the Transformer Network, which sums an input sequence y with a positional encoding vector p (blue). The model encodes (red) the sequence into an encoded representation of the input sequence ω , before decoding (violet) to yielding output sequence $\hat{y}$.

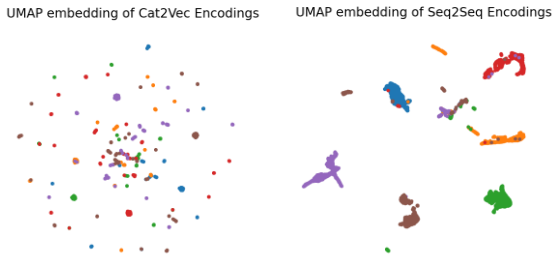

Fig. 5a Fig. 5b

Fig. 5: UMAP visualisation of events encoded from treatment sequences in synthetic treatment data. Points represent treatment events, and colour depicts the treatment group expressed by the event.

Table 1: Adjusted mutual information score of treatment group identification using LDA and our method as |G| and |E| vary.

		LDA (event)			CaSE (event)			CaSE + PHC (event)		
	$\|\mathbb{E}\|$	100	1000	10000	100	1000	10000	100	1000	10000
$\|\mathbb{G}\|$	6	0.655	0.656	0.627	0.803	0.962	0.960	**0.878**	**0.995**	**0.996**
	12	0.707	0.704	0.709	0.771	0.887	0.958	**0.821**	**0.976**	**0.990**
	24	0.749	0.750	0.774	0.705	0.845	0.878	**0.788**	**0.947**	**0.966**
	48	0.763	0.795	0.796	0.655	0.775	0.844	**0.783**	**0.878**	**0.953**

REAL-WORLD DATA: MIMIC-III

The MIMIC-III dataset is a large database comprising de-identified health-related data from the Beth Israel Deaconess Medical Center. We use sequences of ICD-9 diagnosis codes from hospital visits. After cleaning, the dataset contained 46,520 patients and 267,703 diagnosis events; from which $|\mathbb{E}| = 5{,}262$ unique ICD-9 codes were observed.

We learn representations using events from individual patient treatment sequences, where each event contains an ICD-9 code, and the ontological context. **Fig. 6** visualizes the encoded representations using a 2D UMAP embedding. Like in **Fig. 5b**, **Seq2Seq** groups together distinct treatment events based on learned relationships between them as shown in **Fig. 6a**. When colouring events by their position in a treatment sequence, clusters of events express a dominant colour indicating inter-treatment group dynamics, as shown in **Fig. 6b**. These findings demonstrate that CaSE captures the features that are characteristic of healthcare objectives.

CONCLUSION

We introduce CaSE, a generalised method for discovering latent treatment groups in EHRs. CaSE enables event-level identification of topics in sequences, and outperforms traditional topic models. This approach can be used to identify healthcare objectives in EHRs, and critically enhance healthcare management efforts.

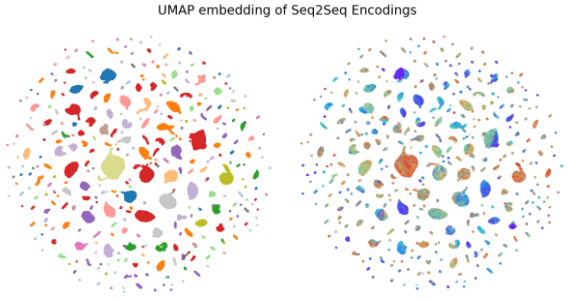

Fig. 6a Fig. 6b

Fig. 6: Fig. 6a depicts encoded events coloured by the ICD-9 ontology. **Fig. 6b** encoded events coloured by the position of each event within the sequence from which it occurred; cool colours indicate events early in the sequence, while warm colours events late in the sequence.

This work was supported by Cancer Australia in the form of a doctoral research stipend (to A. Caruana).