Hierarchical Apprenticeship Learning for Disease Progression Modeling

Xi Yang[1*], **Ge Gao**[2], **Min Chi**[2]
[1] IBM Research
[2] North Carolina State University
xi.yang@ibm.com, {ggao5, mchi}@ncsu.edu

Abstract

Disease progression modeling (DPM) plays an essential role in characterizing patients' historical pathways and predicting their future risks. Apprenticeship learning (AL) aims to induce decision-making policies by observing and imitating expert behaviors. In this paper, we investigate the incorporation of AL-derived patterns into DPM, utilizing a Time-aware Hierarchical EM Energy-based Subsequence (THEMES) AL approach. To the best of our knowledge, this is the first study incorporating AL-derived progressive and interventional patterns for DPM. We evaluate the efficacy of this approach in a challenging task of septic shock early prediction, and our results demonstrate that integrating the AL-derived patterns significantly enhances the performance of DPM.

1 Introduction

Disease progression modeling (DPM) aims to characterize patients' longitudinal medical records, identify disease progressive stages, and evaluate factors affecting the progression pathways [Cook and Bies, 2016]. Empowered by the increasing availability of large medical datasets, *e.g.*, electronic health records (EHRs), various deep learning models [Choi *et al.*, 2016; Zhang *et al.*, 2017] have been developed for DPM based on the recorded *observations* in EHRs, *e.g.*, vital signs and laboratory test results. However, relying solely on monitored observations may not fully capture the complexity of disease progression [Tintinalli *et al.*, 1985]. For example, different patient groups may exhibit varying observations for the same disease, where a normal sign in one group may turn out to be abnormal in another [Baytas *et al.*, 2017]. Some recent DPM works have used *subsequence clustering* for deriving *progressive stages* automatically in a data-driven manner from various observations [Yang *et al.*, 2021]. Other than using observations for DPM, *interventions* also have a significant impact on disease progression [Komorowski *et al.*, 2018; Azizsoltani *et al.*, 2019], and some prior works have taken clinicians' interventions as additional features for DPM [Esteban *et al.*, 2016; Goh *et al.*, 2021]. Clinical interventions are usually carried out based on prior medical knowledge, which is difficult to be reflected by observations alone. In addition, interventions tend to take hours or days to manifest in observations [Dulac-Arnold *et al.*, 2020]. Furthermore, individual clinical interventions on DPM cannot be fully explained without assessing their effectiveness, since different patient groups may respond to the same treatment differently, and patients at different stages of DPM may respond differently to the same treatment [Clifford *et al.*, 2016]. For this reason, when learning progressive patterns in DPM, it is essential to incorporate the *effectiveness of interventions*.

Reinforcement learning (RL) has shown considerable potential in learning effective interventional strategies for clinical decision-making [Wang *et al.*, 2018; Yu *et al.*, 2021]. A common challenge when applying RL to healthcare is the design of *reward function*, which serves as an incentive to induce effective policies [Azizsoltani *et al.*, 2019]. Apprenticeship learning (AL) was proposed to address this issue [Abbeel and Ng, 2004]: instead of taking an explicitly delineated reward function as input, AL learns the policy by imitating the demonstrated behaviors of clinical experts. Recently, an AL method named energy-based distribution matching (EDM), has shown great success in inducing effective clinical interventional policies directly from EHRs [Jarrett *et al.*, 2020]. In EDM, the demonstrations provided by clinicians are assumed to be generated via a uniform policy, latently driven by a *single* reward function. However, while managing patients under various progressive stages, such as common fevers or severe organ failures, clinicians may adopt varying policies with *multiple reward functions* [Wang *et al.*, 2021].

In this paper, we leverage an AL named Time-aware Hierarchical EM Energy-based Subsequence (THEMES) clustering [Yang *et al.*, 2023] for DPM. THEMES is designed to handle the *multiple evolving reward functions* through two key components: a) *subsequence partitioning*, which clusters patients' sequential records into progressive stages, and b) *policy induction*, which induces interventional policy for each stage. The two components are performed iteratively: referring to the *progressive* stages learned by subsequence partitioning, fine-grained evolving interventional policies can be induced; meanwhile, these derived interventional patterns will, in turn, refine the partitioned subsequences to indi-

[*]This research was conducted during the author's affiliation with North Carolina State University.

cate more accurate progressive stages. The two components are modeled by a Reward-regulated Multivariate Time-aware Toeplitz Inverse Covariance-based Clustering (RMT-TICC) and an Expectation Maximization EDM (EM-EDM), respectively. RMT-TICC encodes the *interventional* patterns by introducing a reward regulator during the subsequence partitioning, motivated by the fact that interventions are latently driven by rewards; while EM-EDM captures the *progressive* patterns by using an EM to simultaneously cluster the RMT-TICC learned subsequences and induce their respective policies. We incorporate both *progressive* patterns from RMT-TICC and *interventional* patterns from EM-EDM for DPM.

The effectiveness of incorporating THEMES-learned patterns for DPM is evaluated by modeling *sepsis*, an extremely challenging and life-threatening organ dysfunction that is a leading cause of death worldwide [Singer *et al.*, 2016]. Without timely interventions, patients can progress to the most severe condition of *septic shock* with a high mortality rate of 50% [Martin *et al.*, 2003], while 80% of sepsis deaths can be prevented with timely diagnosis and treatment [Kumar *et al.*, 2006]. Therefore, modeling sepsis progression is crucial for more accurate early prediction of septic shock. To achieve this goal, we leverage THEMES by incorporating latent *progressive* patterns extracted by RMT-TICC and *interventional* patterns derived by EM-EDM as supplemental features for original *observations*. Our results demonstrate that incorporating such additional features leads to improved DPM accuracy. The major contributions of this work are three-fold:

• To our best knowledge, this is the first work incorporating AL-derived *progressive* and *interventional* patterns for DPM.

• Our empirical results reveal that the success of incorporating AL for DPM relies on the fact that THEMES can handle time-awareness and evolving reward functions.

• We utilize AL for modeling a challenging disease, *i.e.*, sepsis, and predicting its most severe condition, *i.e.*, septic shock, in early stages. This indicates the potential to enhance personalized and timely treatments for patients.

2 Related Work

2.1 Disease Progression Modeling

The importance of DPM has been recognized in previous studies [Cook and Bies, 2016; Severson *et al.*, 2020], leading to the usage of deep learning models such as recurrent neural networks [Lipton *et al.*, 2015; Saqib *et al.*, 2018]. Among these models, long short-term memory (LSTM) has shown remarkable success due to its ability to capture sequential patterns from historical records. Typically, these deep learning models take monitored *observations* as input and use neural networks to extract latent progressive patterns for DPM.

An alternative approach to build DPM is through *subsequence clustering*, which captures disease progressive stages by partitioning and clustering subsequences. In general, subsequence clustering can be categorized into distance-based (*e.g.*, dynamic time warping [Giannoula *et al.*, 2018]) and model-based (*e.g.*, Gaussian mixture models [Faruqui *et al.*, 2021] and hidden Markov models [Kwon *et al.*, 2020]). Model-based methods are usually more reliable for better handling noise and outliers. Recently, a model-based approach called Toeplitz inverse covariance-based clustering [Hallac *et al.*, 2017] has gained attention for accurately partitioning subsequences in various applications, such as analyzing physical activities for Alzheimer's patients [Li *et al.*, 2018] and segmenting critical stages for sepsis patients [Gao *et al.*, 2022]. Building upon this work, a multi-series time-aware Toeplitz inverse covariance-based clustering approach (MT-TICC) [Yang *et al.*, 2021] was proposed, which takes multiple series as input and uses a time-awareness mechanism to handle irregular time intervals in EHRs. The MT-TICC-derived progressive patterns, when used as additional information of *observations*, enable more accurate early prediction for septic shock, outperforming competitive baselines.

It is important to recognize that clinicians' *interventions* play a significant role in patient's disease progression [Komorowski *et al.*, 2018; Azizsoltani *et al.*, 2019], yet this is not sufficiently addressed in the majority of existing works. Though some prior studies have included interventions as additional features in DPM [Esteban *et al.*, 2016; Goh *et al.*, 2021], they do not consider the effectiveness of interventions.

2.2 RL & AL for EHRs

Reinforcement learning (RL) has been widely applied in EHRs for dynamic treatment regimes, which aims at inducing a decision-making policy to dictate how the interventions should be executed so that the patients can gain improved outcomes [Yu *et al.*, 2021]. As an input of RL, the reward function plays a critical role in praising/punishing the learning model to derive an optimal policy. However, manually specifying an appropriate reward function is usually expertise-intensive and time-consuming, posing a significant barrier to the broader applicability of RL [Abbeel and Ng, 2004]. **Apprenticeship learning (AL)** [Ng *et al.*, 2000] tackles such problems by directly learning the reward function from experts' demonstrations. *Behavior cloning* [Raza *et al.*, 2012] is a classic AL method, which directly learns a mapping from states to actions to greedily imitate experts' demonstrated behaviors [Ross *et al.*, 2011]. Later, various Inverse RL (IRL) [Abbeel and Ng, 2004; Ziebart *et al.*, 2008] and adversarial imitation learning [Ho and Ermon, 2016; Finn *et al.*, 2016] based AL approaches have been proposed, but they are often *online*, requiring iteratively executing the latest policy to collect data for updating the model. The execution of a bad policy in healthcare is unethical [Levine *et al.*, 2020], making offline AL methods desired. Though some online approaches have been adapted to offline [Chan and van der Schaar, 2021; Kostrikov *et al.*, 2018], they usually rely on off-policy evaluation, which itself is nontrivial with imperfect solutions.

More recently, [Jarrett *et al.*, 2020] introduced an AL approach named energy-based distribution matching (EDM) and evaluated it on various benchmarks, *e.g.*, Acrobot, LunarLander, and BeamRider. Their results demonstrated that EDM outperformed both *IRL-based* and *adversarial imitation learning-based* methods [Brockman *et al.*, 2016]. Therefore, we employ EDM as a baseline in this paper. When applying EDM to EHRs for modeling clinicians' interventions, it makes a strong assumption that all demonstrations are generated with a unified policy, following a *single* re-

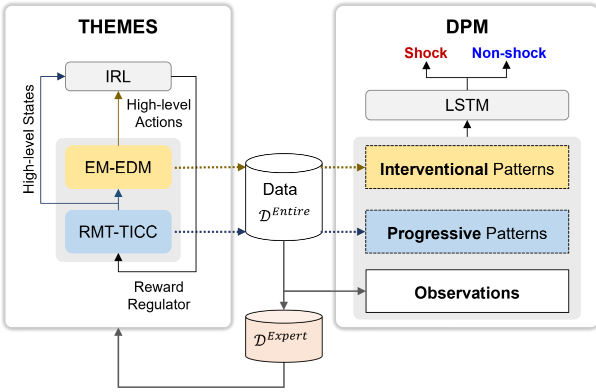

Figure 1: THEMES for DPM.

ward function. However, clinicians commonly execute different policies with multiple reward functions when treating patients under different progressive stages [Wang *et al.*, 2021; Wang *et al.*, 2022]. To handle the *multiple* reward functions, some improved AL have been proposed [Dimitrakakis and Rothkopf, 2011; Babes *et al.*, 2011]. For example, [Babes *et al.*, 2011] developed an EM-based IRL, which assumes reward functions are diverse across different demonstrations, while within each demonstrated sequence, the reward function is assumed to be unified. To model the multiple reward functions *evolving* over time, several more recent works have been proposed [Krishnan *et al.*, 2016; Hausman *et al.*, 2017; Wang *et al.*, 2021; Wang *et al.*, 2022]. However, these methods generally partition the demonstrations into *fixed-length* subsequences to learn their respective reward functions, without considering the irregular intervals during the partitioning.

Furthermore, existing AL methods typically concentrate on learning interventional patterns for either *inducing* a decision-making policy or *evaluating* whether interventions are carried out as expected. However, none of these methods have been incorporated to learn the progressive stages during the DPM.

3 THEMES for DPM

Figure 1 provides an illustration of how THEMES-derived patterns can be utilized for DPM. The entire input data $\mathcal{D}^{Entire}$ consists of L sequences represented as $\{\mathcal{D}^l\} = \{(\mathbf{x}_t^l, a_t^l)|t = 1, ..., T^l; l = 1, ..., L\}$, where $\mathbf{x}_t^l \in \mathbb{R}^m$ is the t-th multivariate observed state with m features, and a_t^l is the corresponding action in the l-th sequence with the length of T^l. In the context of AL, it is typically assumed that the experts' demonstrations provided as input are optimal or near-optimal following a latent reward function [Abbeel and Ng, 2004]. The quality of these demonstrations plays an important role in inducing more accurate policies. Therefore, we have designed a procedure for determining the experts' demonstrations (detailed in Section 4.1), based on which a subset of N ($N \leq L$) sequences $\mathcal{D}^{Expert}$ are selected, denoted as: $\{\mathcal{D}^n\} = \{(\mathbf{x}_t^n, a_t^n)|t = 1, ..., T^n; n = 1, ..., N\}$.

Taking $\mathcal{D}^{Expert}$ as input, THEMES aims at learning the *multiple* underlying reward functions evolving over time. As illustrated in Figure 1 (Left), THEMES has a *hierarchical* structure, with two major components at its low-level, *i.e.*, RMT-TICC and EM-EDM. Specifically, **RMT-TICC** operates on the states in $\mathcal{D}^{Expert}$ as input. It partitions and clusters the subsequences, such that each subsequence cluster exhibits the same time-invariant patterns, which is regarded as a *high-level state*. Then focusing on the state-action pairs over the partitioned subsequences, **EM-EDM** will cluster and induce their policies, such that each cluster exhibits consistent decision-making patterns, which is referred to as a *high-level action*. Subsequently, taking the high-level state-action pairs as input, a *high-level reward regulator* will be learned by IRL and fed back to refine the RMT-TICC. This iterative procedure continues until convergence is achieved.

Once the THEMES model has been learned, we apply it over $\mathcal{D}^{Entire}$ for extracting *progressive* patterns from RMT-TICC and *interventional* patterns from EM-EDM. These patterns are then combined with the *observations* from the original data, which will further serve as an input of LSTM for septic shock early prediction, as depicted in Figure 1 (Right).

3.1 Subsequence Partitioning by RMT-TICC

Preliminaries

Given experts' demonstrations, denoting the states in $\mathcal{D}^{Expert}$ as $\{\mathbf{x}_t^n|t = 1, ..., T^n\}$, the objective of subsequence clustering is partitioning and clustering the $\{\mathbf{x}_t^n\}$ into subsequences based on their latent time-invariant patterns. It can be achieved by learning a mapping from each state to a specific cluster $\{k|k = 1, ..., K\}$. To capture the interdependence among neighboring states, a sliding window of length $\omega \ll T^n$ is employed to incorporate the context information. When assigning a state $\mathbf{x}_t^n$ into a cluster k, we also consider its preceding states within the sliding window, *i.e.*, $\mathbf{X}_t^n = \{\mathbf{x}_{t-\omega+1}^n, ..., \mathbf{x}_t^n\}$, where $\mathbf{X}_t^n$ is a random variable of dimension $m\omega$, concatenating the m-dim states in ω.

To learn the subsequence clusters, we fit $\mathbf{X}_t^n$ into K Gaussian distributions, with each distribution indicating a specific cluster k. It can be modeled by Toeplitz inverse covariance-based clustering [Hallac *et al.*, 2017] to learn the mean and inverse covariance matrix for each cluster. More specifically, the mean vectors, $\{\mu_k|k = 1, ..., K\}$, are determined by assigning each state to an optimal cluster, resulting in clustering assignments $\mathbf{P} = \{P_k|k = 1, ..., K\}$, where $P_k \subset \{1, ..., T\}$ is the indices of states (sliding windows) belonging to cluster k. The inverse covariance matrices, $\mathbf{\Theta} = \{\mathbf{\Theta}_k|k = 1, ..., K\}$, are estimated to characterize the time-invariant structural patterns for each cluster. Herein, $\mathbf{\Theta}_k$ is constrained to be block-wise Toeplitz, composed of ω sub-blocks $A^{(i)} \in \mathbb{R}^{m\times m}$, $i \in [0, \omega - 1]$, where the sub-block $A^{(i)}$ represents the partial correlations among m features between timestamp t and $t+i$. For instance, the (p, q)-th element in $A^{(i)}$ represents the partial correlation between the p-th feature at t and the q-th feature at $t + i$, where $p, q \in \{1, ..., m\}$.

A multi-series time-aware Toeplitz inverse covariance-based clustering (MT-TICC) [Yang *et al.*, 2021] was proposed recently, improving the subsequence clustering from two perspectives: First, MT-TICC operates on *multi-series* data, allowing for improved estimation of the mean and inverse covariance matrix by considering shared patterns across different sequences. This capability is particularly advantageous in real-world scenarios like EHRs, where the data is collected from various patients. Second, MT-TICC incorporates *time-awareness* to handle irregular intervals by using a decay function to constrain the consistency within each cluster. It is highly preferred when dealing with data collected irregularly, such as in healthcare settings. However, when applied to DPM, MT-TICC primarily focuses on the *observed states*, and it does not fully account for the *interventional patterns* conveyed by *state-action pairs*, which can significantly impact the disease progression [Komorowski *et al.*, 2018].

RMT-TICC

To incorporate the *interventional patterns* conveyed by *state-action pairs*, a *Reward-regulated* MT-TICC (RMT-TICC) is leveraged in THEMES, considering the fact that the interventions are driven by underlying *reward functions*. The objective function of RMT-TICC is as follows:

$$\begin{aligned}\underset{\boldsymbol{\Theta},\mathbf{P}}{\text{argmin}} \sum_{k=1}^{K}\Bigg[\sum_{n=1}^{N}\sum_{\mathbf{X}_t^n\in\mathbf{P}_k}\Bigg(&\overbrace{-\ell\ell(\mathbf{X}_t^n,\boldsymbol{\Theta}_k)}^{\text{Log-likelihood}}+\\ &\overbrace{c(\mathbf{X}_{t-1}^n,\mathbf{P}_k,\Delta T_t^n,\Delta r_t^n)}^{\text{Reward-regulated Time-aware Consistency}}\Bigg)+\lambda\overbrace{||\boldsymbol{\Theta}_k||_1}^{\text{Sparsity}}\Bigg]\end{aligned} \tag{1}$$

• *Log-likelihood term* measures the probability that $\mathbf{X}_t^n$ belongs to the cluster k. Specifically, assuming $\mathbf{X}_t^n$ is a Gaussian distribution with the mean of μ_k and the inverse covariance matrix of $\boldsymbol{\Theta}_k^{-1}$, the log-likelihood term is defined as:

$$\begin{aligned}\ell\ell(\mathbf{X}_t^n,\boldsymbol{\Theta}_k)=&-\frac{1}{2}(\mathbf{X}_t^n-\mu_k)^T\boldsymbol{\Theta}_k(\mathbf{X}_t^n-\mu_k)\\ &+\frac{1}{2}\log|\boldsymbol{\Theta}_k|-\frac{m}{2}\log(2\pi)\end{aligned} \tag{2}$$

• *Reward-regulated Time-aware Consistency term* encourages the consecutive states $\{\mathbf{X}_{t-1},\mathbf{X}_t\}$ to be assigned into the same cluster, by taking account of both the time intervals and the corresponding rewards. It penalizes the neighbored states belonging to different clusters by minimizing Eq.(3):

$$c(\mathbf{X}_{t-1}^n,\mathbf{P}_k,\Delta T_t^n,\Delta r_t^n)=\frac{\beta\mathbb{1}\{t-1\notin P_k\}}{\Phi(\Delta r_t^n, log(e+\Delta T_t^n))} \tag{3}$$

In above equation, β is a weight parameter. $\mathbb{1}\{t-1\notin P_k\}$ is an indicator function, with the value of 1 if $\mathbf{X}_{t-1}^n$ and $\mathbf{X}_t^n$ does not belong to the same cluster; otherwise its value is 0. $1/log(e+\Delta T_t^n)$ is a decay function, which adaptively relax the penalization in Eq.(3) when the interval ΔT_t^n between consecutive states becomes larger [Baytas *et al.*, 2017].

To incorporate decision-making patterns for refining the subsequence clustering, a *hierarchical* structure is utilized, as illustrated in Figure 1 (Left). Rather than directly learning rewards from observed state-action pairs, this hierarchical approach is employed for two reasons. First, decision-making patterns across different subsequences have varying degrees of importance contributing the patients outcomes, while treating them equally without considering the hierarchy would overlook their individual significance. Second, learning from state-action pairs in a flattened structure might not fully capture the transitional patterns across different policies.

Based on the high-level states $\{P_k|k=1,..,K\}$ and high-level actions $\{\Pi_g, g=1,...,G\}$, a high-level reward regulator $\overline{R}$ can be learned. It is initialized as $\overline{R}=\mathbf{1}$. In each iteration, we employ a maximum likelihood inverse reinforcement learning (IRL) [Babes *et al.*, 2011] to update the $\overline{R}$, which has demonstrated efficiency in inferring reward functions [Yang *et al.*, 2020]. Based upon $\overline{R}$, the reward $\{r_t^n|t=1,...,T^n\}$ for each state-action pair $(\mathbf{x}_t^n, a_t^n)$ can be calculated as:

$$r_t^n=\frac{1}{G}\sum_{g=1}^{G}\sum_{k=1}^{K}\mathbb{1}\{t\in P_k\}\Pi_g(\mathbf{x}_t^n,a_t^n)\overline{R}(P_k,\Pi_g) \tag{4}$$

In Eq.(4), $\mathbb{1}\{t\in P_k\}$ has the value of 1 if $\mathbf{x}_t^n$ belongs to the subsequence cluster k; otherwise its value is 0. $\Pi_g(\mathbf{x}_t^n,a_t^n)$ denotes the probability of taking a_t^n at $\mathbf{x}_t^n$ with the g-th policy. Based upon the learned reward r_t^n, we calculate the Δr_t^n for consecutive state-action pairs to regulate the consistency constraint. To balance the effects of time-awareness patterns, *i.e.*, $log(e+\Delta T_t^n)$, and decision-making patterns, *i.e.*, Δr_t^n, we employ a bivariate Gaussian distribution Φ to model their interactional regularizations, as shown in Eq.(3).

The introduced reward regulator can capture the interventional patterns and implicitly reflect the effectiveness of treatments. When two consecutive timestamps have similar states but significantly different rewards, it may indicate that the patient's condition has progressed into another stage.

• *Sparsity term* is responsible for controlling the sparseness of the model using an l_1-norm penalty, denoted as $\lambda||\boldsymbol{\Theta}_k||_1$, where λ is a coefficient. This term encourages the selection of the most significant variables, effectively preventing overfitting by reducing the complexity of the model.

To solve Eq.(1), we employed EM to learn the cluster assignments $\mathbf{P}$ and the patterns $\boldsymbol{\Theta}$ iteratively until convergence. Specifically, *in E-step*, we fix $\boldsymbol{\Theta}$ to learn $\mathbf{P}$, then Eq.(1) simplifies to include only the log-likelihood term and the consistency term. It can be solved by dynamic programming to find a minimum cost Viterbi path [Viterbi, 1967]; *In M-step*, by fixing $\mathbf{P}$ to learn the $\boldsymbol{\Theta}$, Eq.(1) reduces to include only the log-likelihood term and the sparsity term. It can be formulated as a typical graphical lasso [Friedman *et al.*, 2008] with a Toeplitz constraint over $\boldsymbol{\Theta}$ and be solved by an alternating direction method of multipliers [Boyd *et al.*, 2011]. We iteratively perform the E-step and M-step until convergence.

3.2 Policy Induction by EM-EDM

Preliminaries

Energy-based distribution matching (EDM) [Jarrett *et al.*, 2020] is a strictly *offline* AL method that learns a policy solely from expert demonstrations $\{\mathcal{D}^n\}$ without requiring knowledge about model transitions or off-policy evaluations. It assumes that the $\{\mathcal{D}^n\}$ are carried out with a policy Π^θ parameterized by θ, driven by a *single* reward function.

For simplicity, we will denote a state-action pair as $(\mathbf{x}, a)$, omitting their indexes when it does not cause ambiguity. The occupancy measures for the demonstrations and for the learned policy are denoted as $\rho_{\mathcal{D}}$ and ρ_{Π^θ}, respectively. The probability density for each state-action pair can be measured as $\rho_{\Pi^\theta}(\mathbf{x},a)=\mathbb{E}_{\Pi^\theta}[\sum_{t=0}^{\infty}\gamma^t\mathbb{1}\{\mathbf{x}_t=\mathbf{x},a_t=a\}]$, where γ is a discount factor. Then, the probability density for each state can be measured by: $\rho_{\Pi^\theta}(\mathbf{x})=\sum_a\rho_{\Pi^\theta}(\mathbf{x},a)$. The goal of inducing the policy Π^θ can be achieved by minimizing the KL divergence between $\rho_{\mathcal{D}}$ and ρ_{Π^θ}:

$$\underset{\theta}{\text{argmin}}\, D_{KL}(\rho_{\mathcal{D}}||\rho_{\Pi^\theta})=\underset{\theta}{\text{argmin}}-\mathbb{E}_{\mathbf{x},a\sim\rho_{\mathcal{D}}}\log\rho_{\Pi^\theta}(\mathbf{x},a) \tag{5}$$

Since $\Pi^\theta(a|\mathbf{x}) = \rho_{\Pi^\theta}(\mathbf{x}, a)/\rho_{\Pi^\theta}(\mathbf{x})$, the objective function can be reformulated as:

$$\underset{\theta}{\operatorname{argmax}} -\mathrm{E}_{\mathbf{x}\sim\rho_{\mathcal{D}}} \log \rho_{\Pi^\theta}(\mathbf{x}) - \mathrm{E}_{\mathbf{x},a\sim\rho_{\mathcal{D}}} \log \Pi^\theta(a|\mathbf{x}) \quad (6)$$

When it is not possible to execute the policy Π^θ in an *online* manner, estimating $\rho_{\Pi^\theta}(\mathbf{x})$ in the first term of Eq.(6) becomes challenging. EDM addresses this issue by employing an energy-based model [Grathwohl *et al.*, 2019].

According to energy-based model, the probability density $\rho_{\Pi^\theta}(\mathbf{x})$ is proportional to $e^{-E(\mathbf{x})}$, where $E(\mathbf{x})$ is an energy function. The occupancy measure for state-action pairs can be represented as $\rho_{\Pi^\theta}(\mathbf{x}, a) = e^{f_{\Pi^\theta}(\mathbf{x})[a]}/Z_{\Pi^\theta}$, and the occupancy measure for states can be obtained by marginalizing out the actions: $\rho_{\Pi^\theta}(\mathbf{x}) = \sum_a e^{f_{\Pi^\theta}(\mathbf{x})[a]}/Z_{\Pi^\theta}$. Herein, Z_{Π^θ} is a partition function, and $f_{\Pi^\theta} : \mathbb{R}^{|X|} \to \mathbb{R}^{|\mathbb{A}|}$ is a parametric function that maps each state to $|\mathbb{A}|$ real-valued numbers.

The parameterization of Π^θ implicitly defines an energy-based model over the states distribution, where the energy function can be defined as: $E_{\Pi^\theta}(\mathbf{x}) = -\log \sum_a e^{f_{\Pi\theta}(\mathbf{x})[a]}$. Within the scope of the energy-based model, the first term in Eq.(6) can be reformulated as an occupancy loss:

$$\mathcal{L}_\rho(\theta) = \mathbb{E}_{\mathbf{x}\sim\rho_{\mathcal{D}}} E_{\Pi^\theta}(\mathbf{x}) - \mathbb{E}_{\mathbf{x}\sim\rho_{\Pi^\theta}} E_{\Pi^\theta}(\mathbf{x}) \quad (7)$$

where $\nabla_\theta \mathcal{L}_\rho(\theta) = -\mathbb{E}_{\mathbf{x}\sim\rho_{\mathcal{D}}} \nabla_\theta \log \rho_{\Pi^\theta}(\mathbf{x})$ can be solved by existing optimizers, *e.g.*, stochastic gradient Langevin dynamics [Welling and Teh, 2011]. Thus, by substituting the first term in Eq.(6) with Eq.(7) via energy-based model, we can derive a surrogate objective function to get the optimal solution without the need for *online* policy rollouts.

EM-EDM

To handle *multiple* reward functions varying across demonstrations, an EM-based inverse reinforcement learning approach was proposed in [Babes *et al.*, 2011]. It iteratively clusters the demonstrations in *E-step* and induces policies for each cluster by IRL in *M-step*. However, in M-step, it relies on IRL methods with *discrete* states, which may not be scalable for large continuous state spaces like EHRs. To handle the *continuous* states, motivated by the EM framework and the success of EDM, THEMES employs an extended method—EM-EDM, with EDM in the M-step of EM.

Taking subsequences $\{\hat{\mathcal{D}}^{\hat{n}}|\hat{n} = 1, ..., \hat{N}\}$ learned by RMT-TICC as input, where $\hat{N}$ represents the number of subsequences, the goal of EM-EDM is to cluster these subsequences and learn cluster-specific policies $\{\Pi_g|g = 1, ...G\}$, with G being the number of clusters. The prior probability for each cluster is denoted as ν_g and the policy parameter is denoted as θ_g. Both ν_g and θ_g are randomly initialized. The objective function of EM-EDM is shown in Eq.(8):

$$\underset{\theta_g}{\operatorname{argmax}}\, \mathcal{L} = \sum_{g=1}^{G}\sum_{\hat{n}=1}^{\hat{N}} log(u_{\hat{n}g}) \quad (8)$$

where $u_{\hat{n}g}$ denotes the probability that subsequence $\hat{\mathcal{D}}^{\hat{n}}$ follows the policy of the g-th cluster. It is defined in Eq.(9), with U being a normalization factor.

$$u_{\hat{n}g} = Pr(\hat{\mathcal{D}}^{\hat{n}}|\theta_g) = \prod_{(\mathbf{x},a)\in\hat{\mathcal{D}}^{\hat{n}}} \frac{\Pi_{\theta_g}(\mathbf{x}, a)\nu_g}{U}, \quad (9)$$

During the EM, in the *E-step*, the probability that subsequence $\hat{\mathcal{D}}^{\hat{n}}$ belonging to the cluster g is calculated by Eq.(9). Then, in the *M-step*, the prior probabilities are updated as $\nu_g = \sum_{\hat{n}} u_{\hat{n}g}/\hat{N}$, and the policy parameters θ_g are learned by EDM. The *E-step* and *M-step* are iteratively executed until convergence. Finally, the output of EM-EDM consists of the clustered subsequences along with their respective policies.

Based on the subsequence clusters learned by RMT-TICC and the decision-making policies learned by EM-EDM in THEMES, we extract progressive patterns and interventional patterns from the entire input data $\mathcal{D}^{Entire}$. For each timestamp $(\mathbf{x}_t, a_t)$, based on Eq.(2), we calculate the probabilities that $\mathbf{x}_t$ belonging to each progressive stage. Additionally, we calculate the probabilities that $(\mathbf{x}_t, a_t)$ follows each policy. These probabilities serve as additional features that are concatenated with the original observations $\mathbf{x}_t$, resulting in an augmented input for the downstream early prediction task. In this paper, we fix the early predictor as LSTM [Zhang *et al.*, 2017] and focus on evaluating the effectiveness of feature engineering. We believe that by combining the patterns derived from THEMES with more advanced prediction models (*e.g.*, [Baytas *et al.*, 2017; Zhang *et al.*, 2019], the performance of early prediction can be further enhanced.

4 Experiments

To assess the effectiveness of THEMES-derived patterns for DPM, we applied it to an EHRs dataset obtained from the Christiana Care Health System. This dataset spans over a period of two and a half years, where each sequence represents a patient's visit consisting of a series of observations and corresponding clinicians' interventions.

4.1 Data Preprocessing

Our sepsis-related study cohort comprised 52,919 visits (sequences) with suspected infection, consisting of 4,224,567 timestamps. We conducted data prepossessing as follows:

• *Feature selection*: We consulted clinicians and selected 14 sepsis progression-related features: 1) *Vital signs*: systolic blood pressure, mean arterial pressure, respiratory rate, oxygen saturation, heart rate, temperature, fraction of inspired Oxygen; and 2) *Lab results*: white blood cell, bilirubin, blood urea nitrogen, lactate, creatinine, platelet, neutrophils.

• *Missingness handling*: We address the missing data by an expert-suggested forward-filling (8 hours for vital signs and 24 hours for lab tests), with the remaining missing values imputed as the mean. This method has shown robustness, especially for septic shock early prediction [Zhang *et al.*, 2019].

• *Tagging septic shock visits*: Clinical labeling commonly relies on diagnosis codes, such as ICD-9, which are primarily intended for administrative and billing purposes, leading to constrained reliability [Zhang *et al.*, 2019]. To address this issue, our clinicians referred to the *Sepsis-3* guidelines [Singer *et al.*, 2016] and established specific rules for identifying septic shock. By combing the ICD-9 codes and clinicians' rules, we identified 1,869 shock and 23,901 non-shock visits. To handle the highly imbalanced ratio, we employed a stratified random sampling technique over the non-shock visits, which

Methods	Hold-off Window $\tau = 12$					Hold-off Window $\tau \in [12, 36]$				
	Acc	Rec	Prec	F1	AUC	Acc	Rec	Prec	F1	AUC
Original	.754(.013)	.737(.012)	.763(.014)	.750(.013)	.827(.014)	.724(.018)	.714(.021)	.731(.020)	.721(.033)	.812(.026)
MT-TICC	.803(.010)	.802(.012)	.802(.011)	.802(.012)	.861(.013)	.776(.020)	.790(.024)	.771(.019)	.778(.021)	.839(.020)
Action	.757(.013)	.754(.013)	.759(.012)	.756(.013)	.832(.014)	.724(.018)	.717(.020)	.723(.019)	.722(.029)	.814(.024)
EDM	.765(.013)	.753(.012)	.772(.013)	.762(.011)	.821(.013)	.727(.017)	.718(.020)	.736(.019)	.726(.031)	.829(.024)
THEMES_0	**.809(.011)**	**.837(.012)***	**.793(.012)**	**.814(.013)**	**.871(.012)**	**.784(.016)**	**.807(.021)**	**.783(.018)**	**.795(.017)**	**.850(.019)**
THEMES	**.834(.011)***	**.820(.012)**	**.843(.012)***	**.832(.011) ***	**.891(.012)***	**.801(.015)***	**.816(.021)***	**.789(.019)***	**.802(.016)***	**.860(.017)***

Table 1: Early prediction in EHRs. The best methods are in bold with *, and the second-best is in bold only.

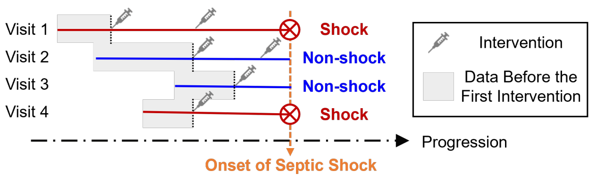

Figure 2: Determining experts' demonstrations by comparing early prediction *before the first intervention* vs. the *onset of septic shock*.

ensured that the dataset maintained the original distribution of age, sex, ethnicity, and stay duration, while achieving a balanced amount. Thus, the non-shock visits were refined to include 1,869 visits, equaling to the number of shock visits.

• *States & Actions*: a) *States* are defined based on the 14 continuous features that are relevant to the progression of sepsis. In addition to the original features, we calculate the maximum and minimum values observed within the past 1 hour for each feature. This allows us to capture temporal changes and trends in the data. As a result, the states are represented as 42-dimensional vectors. b) *Actions* are binary values indicating whether specific antibiotics (*e.g.*, clindamycin, daptomycin) have been administered or not. It is well established, as suggested in [Gauer, 2013], that antibiotic therapy plays a crucial role in enhancing the clinical outcomes of sepsis patients.

• *Determining Experts' Demonstrations*: To enhance the selection of high-quality demonstrations for inducing more accurate policies, we developed a strategy to evaluate the effectiveness of interventions. This strategy involved comparing the early prediction of septic shock *before the first intervention* to the actual outcome at the onset of septic shock, as depicted in Figure 2. To assess the effectiveness of interventions, we trained multiple LSTMs (100 times) using different hyperparameter settings over the data prior to the first intervention for randomly selected visits. For a given visit, if more than 80% of the LSTMs produce the same early prediction result, we compare this prediction with the actual onset of septic shock: If a patient transitions from shock (prior to the first intervention) to non-shock (onset of septic shock), it indicates the interventions were effective, suggesting the visit to be treated an experts' demonstration for policy induction by THEMES; Otherwise, the visit would not be included during the THEMES modeling process. Using this approach, we identified 195 sequences as experts' demonstrations.

4.2 Experimental Settings

We compared the following methods for evaluating the effectiveness of progressive and interventional patterns in DPM:

• *Original:* Except for the *Original* observed state features, none of other additional information will be incorporated;

• *MT-TICC:* Observed state features are supplemented with additional progressive patterns learned by *MT-TICC* [Yang *et al.*, 2021]. The additional features are extracted as the probabilities belonging to each subsequence cluster.

• *Action:* Interventional *Actions* are directly taken as additional features for the observed state features.

• *EDM:* An additional interventional feature is extracted by *EDM*-learned policy [Jarrett *et al.*, 2020], represented as the probability of following the learned policy.

• *THEMES_0:* Additional progressive and interventional patterns are derived from the probabilities associated with the subsequence clusters learned by *MT-TICC* and the probabilities adherent to the policies learned by *EM-EDM*. It is a simplified version of THEMES, where the interventional patterns will not feed back to refine the subsequence clustering.

• *THEMES:* Based on THEMES, the probabilities belonging to *RMT-TICC*-learned subsequence clusters and the probabilities following *EM-EDM*-learned policies are respectively taken as additional progressive and interventional features.

As demonstrated in prior works, EDM outperformed competitive AL methods with a single reward function [Jarrett *et al.*, 2020]. Meanwhile, THEMES_0 and THEMES could induce more accurate policies compared to competitive AL baselines with multiple reward functions (*e.g.*, hierarchical IRL [Krishnan *et al.*, 2016] and multi-modal imitation learning [Hausman *et al.*, 2017]) as well as their ablations (*e.g.*, EM-EDM, and learning progressive stages by MT-TICC then inducing cluster-wise policy by EDM) [Yang *et al.*, 2023]. Therefore, in this paper, we employed EDM, THEMES_0, and TMEMES for extracting AL-derived patterns for DPM.

We utilized Keras to implement the LSTM and performed parameter tuning through grid search. The results were measured when: $\tau = 12$ and $\tau \in [12, 36]$, for evaluating if septic shock prediction could be achieved 12 hours prior to the onset or even earlier before 36 hours. Each method was repeated 10 times, with the data randomly divided into 80% for training and 20% for testing. The evaluation metrics include Accuracy (Acc), Recall (Rec), Precision (Prec), F1-score (F1), and AUC. All parameters in THEMES are determined by 5-fold cross-validation. In RMT-TICC, the cluster number K is set to 11 based on Bayesian information criteria (BIC) [Friedman *et al.*, 2001], the window size ω is set to 2, and the sparsity λ and consistency β coefficients are set to 1e-5 and 4, respectively. In EM-EDM, the cluster number is determined heuristically as 3, by iteratively applying the EM algorithm until

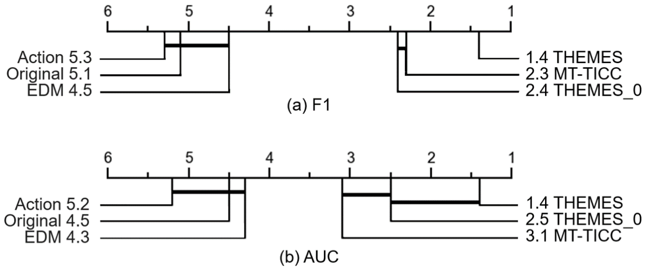

Figure 3: Critical difference diagram with Wilcoxon signed-rank test over (a) F1 and (b) AUC.

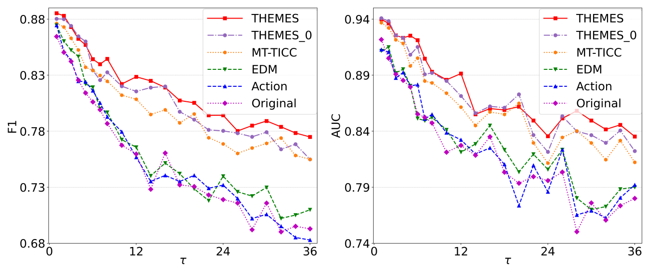

Figure 4: F1 & AUC for septic shock early prediction ($\tau \in [1, 36]$) with additional features learned by different methods.

empty clusters are generated or the log-likelihood varies less than a predefined threshold. The THEMES approach uses a threshold of 10 iterations, as our observed that the clustering likelihood for both MT-TICC and EM-EDM converges within 10 iterations. To ensure fair comparisons, optimal parameters for other baselines are also determined by cross-validation.

4.3 Results

The experimental results are reported in Table 1. The best results are in bold highlighted with * and the second-best results are in bold only. Additionally, we provide Critical Difference diagrams with Wilcoxon signed-rank tests over F1 and AUC in Figure 3. In the diagrams, unconnected models indicate pairwise significance at a confidence level of 0.05.

According to the results: a) *For progressive patterns*: Comparing MT-TICC to Original, the incorporation of progressive patterns in MT-TICC yields significant improvements. Moreover, among the four methods incorporating interventions (*i.e.*, Action, EDM, THEMES_0, and THEMES), the two with progressive patterns (*i.e.*, THEMES_0 and THEMES) outperform the others (*i.e.*, Action and EDM). As a result, the inclusion of progressive patterns enables a better capture of progressive stages, leading to improved early prediction. b) *For interventional patterns*: Comparing Action and EDM vs. Original, the Action yields similar performance to Original, while EDM shows slightly better performance. Additionally, comparing THEMES_0 and THEMES vs. MT-TICC, THEMES_0 is marginally better than MT-TICC, while THEMES further improves the performance. Thus, directly taking actions as additional features via Action cannot fully capture the clinicians' decision-making patterns, while incorporating AL-learned patterns can better reflect the effectiveness of interventions; c) *For both progressive and interventional patterns*: THEMES performs better than THEMES_0, which indicates the effectiveness of using interventional patterns to refine the learned progressive stages.

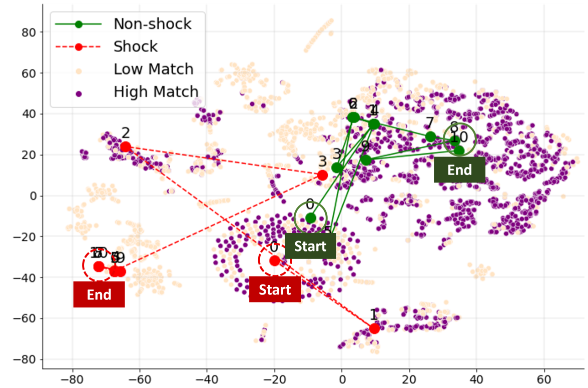

Figure 5: Visualizing the progression of patients.

Figure 4 shows the F1 and AUC when $\tau \in [1, 36]$ hours before the onset of septic shock. As τ increases, it becomes harder to early predict across all models. The figures show the advantage of incorporating progressive patterns, since MT-TICC, THEMES_0, and THEMES outperform Original, Action, and EDM. Meanwhile, incorporating interventions is important, and how they are incorporated also matters, as Action and EDM get similar results with Original, while THEMES_0 and THEMES significantly improve the results.

Figure 5 demonstrates the projection of THEMES-derived features for 100 patients whose interventions best match the learned policies (High Match) and 100 patients with the lowest matching rate (Low Match) on a 2D scatter plot utilizing t-SNE. One shock (red) and one non-shock (green) patient with the same length are randomly sampled. Though the start points for the two patients are close, they drift apart as the non-shock patient's treatments (green) align with the High Match group, while the shock patient's treatments (red) align with the Low Match group. It demonstrates that THEMES can capture the effectiveness of treatments for progression to distinguish shock patients from non-shock patients.

5 Conclusions

In this paper, we explore incorporating AL-derived patterns for DPM. Building upon the success of subsequence clustering in extracting progressive patterns for DPM, we leverage an AL approach named THEMES to capture both progressive and interventional patterns. Taking advantage of these patterns, we aim to handle a challenging task for septic shock early prediction. The experimental results demonstrate that the inclusion of THEMES-derived patterns leads to improved accuracy in predicting septic shock at earlier stages. This advancement holds significant potential for assisting clinicians in delivering timely and personalized treatments.

Acknowledgements

This research was supported by the NSF Grants: #2013502, #1726550, #1660878, and #1651909.

References

[Abbeel and Ng, 2004] Pieter Abbeel and Andrew Y Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the 21st international conference on Machine learning*, page 1. ACM, 2004.

[Azizsoltani *et al.*, 2019] Hamoon Azizsoltani, Yeojin Kim, Markel Sanz Ausin, Tiffany Barnes, and Min Chi. Unobserved is not equal to non-existent: Using gaussian processes to infer immediate rewards across contexts. In *In Proceedings of the 28th IJCAI*, 2019.

[Babes *et al.*, 2011] Monica Babes, Vukosi Marivate, Kaushik Subramanian, and Michael L Littman. Apprenticeship learning about multiple intentions. In *Proceedings of the 28th ICML*, pages 897–904, 2011.

[Baytas *et al.*, 2017] Inci M Baytas, Cao Xiao, Xi Zhang, Fei Wang, Anil K Jain, and Jiayu Zhou. Paient subtyping via time-aware lstm networks. In *SIGKDD*. ACM, 2017.

[Boyd *et al.*, 2011] Stephen Boyd, Neal Parikh, and Eric Chu. *Distributed optimization and statistical learning via the alternating direction method of multipliers*. Now Publishers Inc, 2011.

[Brockman *et al.*, 2016] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

[Chan and van der Schaar, 2021] Alex J Chan and Mihaela van der Schaar. Scalable bayesian inverse reinforcement learning. *arXiv preprint arXiv:2102.06483*, 2021.

[Choi *et al.*, 2016] 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 *NIPS*, pages 3504–3512, 2016.

[Clifford *et al.*, 2016] Kalin M Clifford, Eliza A Dy-Boarman, Krystal K Haase, Kristen Maxvill, Steven E Pass, and Carlos A Alvarez. Challenges with diagnosing and managing sepsis in older adults. *Expert review of anti-infective therapy*, 14(2):231–241, 2016.

[Cook and Bies, 2016] Sarah F Cook and Robert R Bies. Disease progression modeling: key concepts and recent developments. *Current pharmacology reports*, 2016.

[Dimitrakakis and Rothkopf, 2011] Christos Dimitrakakis and Constantin A Rothkopf. Bayesian multitask inverse reinforcement learning. In *European workshop on reinforcement learning*, pages 273–284. Springer, 2011.

[Dulac-Arnold *et al.*, 2020] Gabriel Dulac-Arnold, Nir Levine, Daniel J Mankowitz, Jerry Li, Cosmin Paduraru, Sven Gowal, and Todd Hester. An empirical investigation of the challenges of real-world reinforcement learning. *arXiv preprint arXiv:2003.11881*, 2020.

[Esteban *et al.*, 2016] Cristóbal Esteban, Oliver Staeck, Stephan Baier, Yinchong Yang, and Volker Tresp. Predicting clinical events by combining static and dynamic information using recurrent neural networks. In *2016 IEEE ICHI*, pages 93–101. Ieee, 2016.

[Faruqui *et al.*, 2021] Syed Hasib Akhter Faruqui, Adel Alaeddini, Jing Wang, Carlos A Jaramillo, and Mary Jo Pugh. A functional model for structure learning and parameter estimation in continuous time bayesian network: An application in identifying patterns of multiple chronic conditions. *IEEE Access*, 2021.

[Finn *et al.*, 2016] Chelsea Finn, Sergey Levine, and Pieter Abbeel. Guided cost learning: Deep inverse optimal control via policy optimization. In *International conference on machine learning*. PMLR, 2016.

[Friedman *et al.*, 2001] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

[Friedman *et al.*, 2008] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 2008.

[Gao *et al.*, 2022] Ge Gao, Qitong Gao, Xi Yang, Miroslav Pajic, and Min Chi. A reinforcement learning-informed pattern mining framework for multivariate time series classification. In *IJCAI*, 2022.

[Gauer, 2013] Robert Gauer. Early recognition and management of sepsis in adults: the first six hours. *American family physician*, 88(1):44–53, 2013.

[Giannoula *et al.*, 2018] Alexia Giannoula, Alba Gutierrez-Sacristán, Álex Bravo, Ferran Sanz, and Laura I Furlong. Identifying temporal patterns in patient disease trajectories using dynamic time warping: a population-based study. *Scientific reports*, 8(1):1–14, 2018.

[Goh *et al.*, 2021] Kim Huat Goh, Le Wang, Adrian Yong Kwang Yeow, et al. Artificial intelligence in sepsis early prediction and diagnosis using unstructured data in healthcare. *Nature communications*, 12(1):1–10, 2021.

[Grathwohl *et al.*, 2019] Will Grathwohl, Kuan-Chieh Wang, Jörn-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. Your classifier is secretly an energy based model and you should treat it like one. *arXiv preprint arXiv:1912.03263*, 2019.

[Hallac *et al.*, 2017] David Hallac, Sagar Vare, Stephen Boyd, and Jure Leskovec. Toeplitz inverse covariance-based clustering of multivariate time series data. In *SIGKDD*, pages 215–223, 2017.

[Hausman *et al.*, 2017] Karol Hausman, Yevgen Chebotar, Stefan Schaal, Gaurav Sukhatme, and Joseph J Lim. Multi-modal imitation learning from unstructured demonstrations using generative adversarial nets. *Advances in neural information processing systems*, 30, 2017.

[Ho and Ermon, 2016] Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. *Advances in neural information processing systems*, 29, 2016.

[Jarrett *et al.*, 2020] Daniel Jarrett, Ioana Bica, and Mihaela van der Schaar. Strictly batch imitation learning by energy-based distribution matching. *Advances in Neural Information Processing Systems*, 2020.

[Komorowski *et al.*, 2018] Matthieu Komorowski, Leo A Celi, et al. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature medicine*, 24(11):1716–1720, 2018.

[Kostrikov *et al.*, 2018] Ilya Kostrikov, Kumar Krishna Agrawal, Debidatta Dwibedi, Sergey Levine, and Jonathan Tompson. Discriminator-actor-critic: Addressing sample inefficiency and reward bias in adversarial imitation learning. *ICLR*, 2018.

[Krishnan *et al.*, 2016] Sanjay Krishnan, Animesh Garg, Richard Liaw, Lauren Miller, Florian T Pokorny, and Ken Goldberg. Hirl: Hierarchical inverse reinforcement learning for long-horizon tasks with delayed rewards. *arXiv preprint arXiv:1604.06508*, 2016.

[Kumar *et al.*, 2006] Anand Kumar, Daniel Roberts, et al. Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock. *Critical care medicine*, 2006.

[Kwon *et al.*, 2020] Bum Chul Kwon, Peter Achenbach, Jessica L Dunne, et al. Modeling disease progression trajectories from longitudinal observational data. In *AMIA Annual Symposium Proceedings*, 2020.

[Levine *et al.*, 2020] Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

[Li *et al.*, 2018] Jia Li, Yu Rong, Helen Meng, et al. Tatc: Predicting alzheimer's disease with actigraphy data. In *Proceedings of the 24th ACM SIGKDD*, 2018.

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

[Martin *et al.*, 2003] Greg S Martin, David M Mannino, Stephanie Eaton, and Marc Moss. The epidemiology of sepsis in the united states from 1979 through 2000. *New England Journal of Medicine*, 348(16):1546–1554, 2003.

[Ng *et al.*, 2000] Andrew Y Ng, Stuart Russell, et al. Algorithms for inverse reinforcement learning. In *ICML*, 2000.

[Raza *et al.*, 2012] Saleha Raza, Sajjad Haider, and Mary-Anne Williams. Teaching coordinated strategies to soccer robots via imitation. In *2012 IEEE ROBIO*, 2012.

[Ross *et al.*, 2011] Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the 14th international conference on artificial intelligence and statistics*, pages 627–635, 2011.

[Saqib *et al.*, 2018] Mohammed Saqib, Ying Sha, and May D Wang. Early prediction of sepsis in emr records using traditional ml techniques and deep learning lstm networks. In *40th IEEE EMBC*, pages 4038–4041, 2018.

[Severson *et al.*, 2020] Kristen A Severson, Lana M Chahine, Luba Smolensky, Kenney Ng, Jianying Hu, and Soumya Ghosh. Personalized input-output hidden markov models for disease progression modeling. In *Machine Learning for Healthcare Conference*. PMLR, 2020.

[Singer *et al.*, 2016] Mervyn Singer, Clifford S Deutschman, Christopher Warren Seymour, et al. The third international consensus definitions for sepsis and septic shock (sepsis-3). *Jama*, 315(8):801–810, 2016.

[Tintinalli *et al.*, 1985] Judith E Tintinalli, Gabor D Kelen, J Stephan Stapczynski, et al. *Emergency medicine: a comprehensive study guide*. Mcgraw-hill New York, 1985.

[Viterbi, 1967] Andrew Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE transactions on Information Theory*, 1967.

[Wang *et al.*, 2018] Lu Wang, Wei Zhang, Xiaofeng He, and Hongyuan Zha. Supervised reinforcement learning with recurrent neural network for dynamic treatment recommendation. In *Proceedings of 24th SIGKDD*, 2018.

[Wang *et al.*, 2021] Lu Wang, Wenchao Yu, Wei Cheng, Bo Zong, and Haifeng Chen. Hierarchical imitation learning with contextual bandits for dynamic treatment regimes. In *RL4RealLife Workshop in the 38 th ICML*, 2021.

[Wang *et al.*, 2022] Lu Wang, Ruiming Tang, Xiaofeng He, and Xiuqiang He. Hierarchical imitation learning via subgoal representation learning for dynamic treatment recommendation. In *Proceedings 15th ACM ICWSDM*, 2022.

[Welling and Teh, 2011] Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th ICML*, 2011.

[Yang *et al.*, 2020] Xi Yang, Guojing Zhou, Michelle Taub, Roger Azevedo, and Min Chi. Student subtyping via em-inverse reinforcement learning. *International Educational Data Mining Society*, 2020.

[Yang *et al.*, 2021] Xi Yang, Yuan Zhang, and Min Chi. Multi-series time-aware sequence partitioning for disease progression modeling. In *IJCAI*, 2021.

[Yang *et al.*, 2023] Xi Yang, Ge Gao, and Min Chi. An offline time-aware apprenticeship learning framework for evolving reward functions. *arXiv preprint arXiv:2305.09070*, 2023.

[Yu *et al.*, 2021] Chao Yu, Jiming Liu, Shamim Nemati, and Guosheng Yin. Reinforcement learning in healthcare: A survey. *ACM Computing Surveys (CSUR)*, 2021.

[Zhang *et al.*, 2017] Yuan Zhang, Chen Lin, Min Chi, Julie Ivy, Muge Capan, and Jeanne M Huddleston. Lstm for septic shock: Adding unreliable labels to reliable predictions. In *2017 IEEE Big Data*, pages 1233–1242, 2017.

[Zhang *et al.*, 2019] Yuan Zhang, Xi Yang, Ivy Julie, and Min Chi. Attain: Attention-based time-aware lstm networks for disease progression modeling. In *In Proceedings of the 28th IJCAI*, 2019.

[Ziebart *et al.*, 2008] Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, and Anind K Dey. Maximum entropy inverse reinforcement learning. In *AAAI*, volume 8, pages 1433–1438. Chicago, IL, USA, 2008.