EVA : Generating Longitudinal Electronic Health Records Using Conditional Variational Auto Encoders

DA 2021312101 문구영

Abstract

문제 의식

- 헬스 케어 분야에서 AI를 활용한 solution을 제공하기 위해 현실의 Longitudinal EHR 데이터가 필요
  ↔ 데이터 보안 & 개인 정보 침해
- 재식별화의 위험 & 데이터의 양 부족

목적

- **'현실적인'** EHR Sequence의 생성 (불연속적인 clinical visits + 동일 변수들에 대한 관찰)
  → 환자들간의 차이를 설명할 수 있고, 연구 대상이 되는 질병에 대해 특정화 가능

  ex) 암에 걸린 환자들에 대한 sequence 데이터를 생성
  (conditioned on specific disease conditions, thus enabling disease specific studies)

Abstract

방법론

- **SGD Markov Chain Monte Carlo** : 랜덤 표본을 추출하여 함수의 값을 확률적으로 계산
- **Variational Inference** : 계산이 어려운 확률 분포 P를 다루기 쉬운 확률 분포 q로 근사

  * Monte Carlo 방법을 이용하면 q를 어떤 분포든 사용할 수 있음 (VAE : 정규 분포)

결론

- **현실적인** EHR Sequence 합성 데이터를 생성할 수 있다.
- 실제 EHR 데이터로 훈련된 모델의 예측 성능 ≈ 생성된 EHR 데이터로 훈련된 모델의 예측성능
- **생성된 합성 데이터를 활용하여 실제 데이터를 증식했을 때 최적의 예측 성능을 보인다.**

1. Introduction

합성 EHR 데이터의 필요성

- 재식별화가 불가능, 완벽한 보안
- 연구 목적에 적합한 현실적이고 (실제 Longitudinal HER의 유의미한 패턴 보존), 대용량의 EHR 데이터 제공

기존 생성 데이터 연구의 한계

- Sequence 데이터 생성에 있어 한계를 보임 (only generating a static patient representation without temporal variation)
- 영상, 음성 같은 **연속적 sequence** 데이터 생성은 뛰어난 발전을 보인 반면, **불연속적인 sequence** 데이터 생성은 부진

→ **EHR Variational Autoencoder : generating realistic discrete EHR code sequences**

본 연구의 개선점

① 관심 대상이 되는 의학적 조건에 특정적인 **specific sequence** 생성 가능 – conditional generation

② Sequence의 **다양성** 보장 - parameter의 불확실성

③ 보다 효율적인 사후 확률 분포의 추론 가능 - SGMCMC + Variational Inference

④ 생성된 EHR sequence 데이터의 유용성을 검증 – 모델의 예측성능

2. Generative Models for EHRs

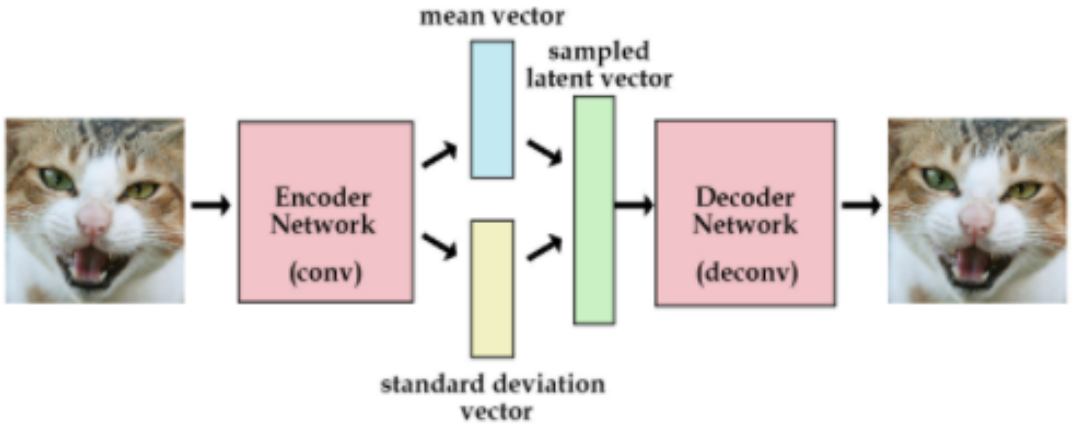
개별 환자의 표현 (가변 길이의 sequence) $x_{n,1:T_n} = \{x_{n,1}, x_{n,2}, \ldots, x_{n,T_n}\}$

- n번째 환자가 병원에 T_n번 방문 (총 T_n개의 record로 구성된 sequence)
- $x_{n,t}$: 고정된 V 차원의 이진 벡터 $x_{n,t}[v] = 1 \; if \; v^{th} \; code \; for \; patient \; n \; was \; observed \; at \; t$

Sequence 데이터의 특징

- records간의 **자기 상관성** (시점에 의존적) ← Autoregressive Likelihoods

VAE



- 입력을 2개의 벡터로 매핑, 매핑된 벡터는 **잠재 공간의 확률 분포를 정의**
- 입력변수의 분포와 이 **분포에 포함된 모수** 추정

2. Generative Models for EHRs

- 입력 sequence x의 분포 $\boldsymbol{p_\theta(x)}$를 추정 (확률 분포 학습) → variation이 있는 sequence 생성 가능
- **입력 sequence의 패턴, 특징을 내재한 잠재변수 z 의 확률 분포의 모수**

Encoder : 입력을 받아 잠재변수 z를 반환 $\boldsymbol{p_\theta(z|x)}$ / z의 평균과 분산

Decoder : encoder가 만들어낸 z의 평균과 분산을 모수로 하는 정규 분포

$$\boldsymbol{p_\theta(x|f_\theta(z)) = N(x|f_\mu(z), f_\sigma(z))}$$

$\boldsymbol{f}$: neural network - z를 활용해 x 복원

$\boldsymbol{\theta}$: 모든 환자들에게 공유되는 global parameter (↔ z : patient specific)

Objective : $\boldsymbol{Maximize\ log\ p_\theta(x) = log\ \sum_z p_\theta\left(x\,|f_\mu(z), f_\sigma(z)\right)}$

→ 잠재 벡터의 모수가 주어졌을 때 x의 확률 분포를 최대한 잘 복원할 수 있도록

- **Variational Inference** 이용, True posterior $p(z_n|x_n;\theta) \approx Tractable\ surrogate\ q(z_n|x_n\ ;\ \emptyset)$
- q가 p를 얼마나 잘 근사하느냐에 따라 VAE의 성능이 결정됨

Variational Inference from Decoder Perspective

$$p(\mathcal{D};\theta) \geq \mathcal{L}(\theta,\phi) \quad ①$$

$$= \sum_n \mathbb{E}_{q(z_n|x_n;\phi)}[\ln p(x_n \mid z_n;\theta)] - \mathrm{KL}(q(z_n \mid x_n;\phi) \,||\, p(z_n)),$$

② ③

- D : $Total\ Sequence\ \{x_1, \dots x_n\}$
- $P(D;\theta)$: 최적화 대상 (어려운 문제)
- $L(\theta, \emptyset)$: 근사 분포 q의 $likelihood$

① Maximize $p(D;\theta)$ ↔ Maximize **lower bound** of $p(D;\theta)$ by using q (근사 분포를 이용해 복잡한 문제를 쉽게 해결)

② $\mathrm{Encode}r$에 의해 z가 주어졌을 때 x에 대한 **복원 오차**

③ 근사 분포 $\boldsymbol{q}$가 참 분포 $\boldsymbol{p}$와 유사해야 한다는 조건 * KL : 두 확률 분포의 유사도 $(KL = 0\ if\ \ q = p)$

Sequential data

- 잠재 벡터 z를 이용해 조건부 확률 분포 $\boldsymbol{p}(\boldsymbol{x}_{\boldsymbol{n},\boldsymbol{1}:\boldsymbol{T}_{\boldsymbol{n}}} \mid \boldsymbol{z}_{\boldsymbol{n}}; \boldsymbol{\theta})$ 를 생성
- Autoregressive model 이용 (시점에 의존적인 패턴 반영)

Sequence wide latent variable

$$\prod_t P(x_{n,t} \mid x_{n,1} \cdots x_{n,t-1}, z_n; \theta)$$

$$= P(x_{n,1}) \cdot P(x_{n,2} \mid x_{n,1}, z_n; \theta) \cdot P(x_{n,3} \mid x_{n,2}, x_{n,1}, z_n; \theta)$$

$$\cdots \cdot P(x_{n,t} \mid x_{n,1}, \cdots, x_{n,t-1}, z_n; \theta)$$

* 훈련된 모델로 문장을 생성하는 과정으로도 이해할 수 있음

2.1 EVA (Electronic health record Variational Auto-encoders)

개별 instance sequence 생성

$$z_n \sim \mathcal{N}(0, \mathbf{I}); \quad x_{n,1:T_n} \mid z_n, \theta \sim p(x_{n,1:T_n} \mid f_\theta(z_n)),$$

①

$$p(x_{n,1:T_n} \mid f_\theta(z_n)) = p(x_{n,1} \mid \xi(f_\theta(z_n))) \prod_{t=2}^{T_n} p(x_{n,t} \mid \xi(f_\theta(x_{n,t-1}, \ldots, x_{n,t-s}, z_n))),$$

②

* 과거의 모든 시점이 아닌, 특정 s 이전 시점까지의 영향을 받음

① 정규 분포를 따르는 잠재 벡터 z

② $neural\ network\ f$는 압축된 표현 z를 sequence 길이에 맞게 펼침 (up-sampling by deconvolution)

③ Sequential dependency는 펼쳐진 z에 1-D 합성곱을 적용시켜 모델링 (Dilation & Masking)

최종 sequence 생성

$$p(\mathcal{D}, \{z_n\}_{n=1}^N, \theta) = p(\theta) \prod_{n=1}^{N} p(z_n) p(x_{n,1} \mid \xi(f_\theta(z_n))) \prod_{t=2}^{T_n} p(x_{n,t} \mid \xi(f_\theta(x_{n,t-1}, \ldots, x_{n,t-s}, z_n))).$$

- 학습된 z, θ가 주어졌을 때 sequence 생성
- Global variable θ, Patient specific z 를 이용하는데 맥락까지(size=s) 고려하여 개별 환자의 sequence 생성

2.2 EVA_c (Hierarchically Factorized Conditional)

EVA의 단점

- EVA는 sequence의 **통제된 생성**이 불가능 (noise sampling)
  → 연구자들의 관심 대상이 되는 질병을 겪고 있는 환자들의 **conditioned sequence**를 생성하는 것이 중요
- 생성 프로세스를 세밀하게 조정할 수 없음
  → 같은 질병을 다른 강도로(severity) 겪는 환자들의 sequence를 생성하는 것이 불가능

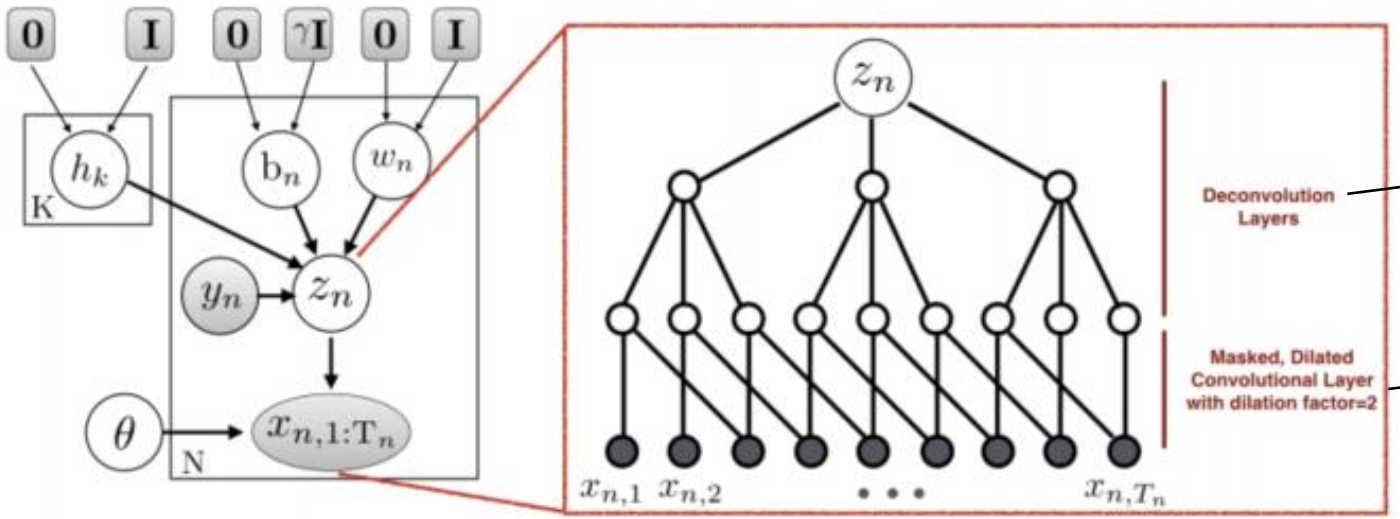
EVA_c

- 환자들의 **meta data** 이용(e.g 해당 환자가 심장병을 앓고 있는 지) → $\boldsymbol{y_{n,k}} = 1 \ if \ patient \ n \ was \ diagnosed \ with \ condition \ k$
- 모든 환자들에게 공유되는 총 K개의 $\boldsymbol{condition}$ → $\boldsymbol{H} = [\boldsymbol{h_1}, \ldots \boldsymbol{h_k}] \ \ where \ \ h_i \sim N(0, I_D), \ D = \dim(z)$
- $patient \ specific \ \boldsymbol{z_n} = span(h_1, \ldots h_k) = w_1 h_1 + \ \ldots + w_k h_k$
- $patient \ specific \ \boldsymbol{w_n} \sim N(0, I_k)$ → 개별 환자가 겪고 있는 condition에 대한 가중치 (severity)
- $patient \ specific \ biases \ \boldsymbol{b_n} \sim N(0, \ \gamma I_K)$

2.2 EVA_c (Hierarchically Factorized Conditional)

$$z_n = \mathrm{H}\pi_n + b_n + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \tau \mathrm{I}_D), \quad \pi_n = y_n \odot \sigma(w_n).$$

- $\boldsymbol{\sigma}$: $logistic$
- $\boldsymbol{\pi_{n,k}}$: condition k에 대한 환자 n의 가중치 (intensity with which patients n expresses condition k)



up-sampling (압축된 표현을 펼침)

시계열 패턴(맥락)을 반영한 sequence 생성

→ global parameter θ를 고려하여 잠재 벡터 z를 EHR sequence로 변환

2.2 EVA_c (Hierarchically Factorized Conditional)

EVA_c Model

$$p(\mathcal{D}, \mathcal{Z}, \theta, \mathbf{H} \mid \eta) = p(\theta)p(\mathbf{H}) \prod_{n=1}^{N} p(w_n)p(b_n \mid \gamma) \quad \eta = \{\{y_n\}_{n=1}^{N}, \tau, \gamma\} \text{ and } \mathcal{Z} = \{z_n, w_n, b_n\}_{n=1}^{N}.$$
$$p(z_n \mid \mathbf{H}, \pi_n, b_n, \tau)p(x_{n,1:T_n} \mid f_\theta(z_n)),$$

∗ $\boldsymbol{\tau}, \boldsymbol{\gamma}$: sequence의 랜덤성에 대한 하이퍼 파라미터, 작게 설정할 수록 condition specific한 sequence 생성

① 환자의 condition y가 주어지면, 그 condition에 따른 **H의 공통된 분포**를 고려한다

② **개별 환자의 고유한 가중치와 편향**을 고려하는데, 그 편향과 가중치, scale parameter에 의해 **개별 환자의 고유한 잠재 벡터**가 주어진다

③ 개별 환자의 고유한 잠재 벡터가 주어지면 **neural network f** 를 이용하여 시계열 sequence로 변환한다

2.2 EVA_c (Hierarchically Factorized Conditional)

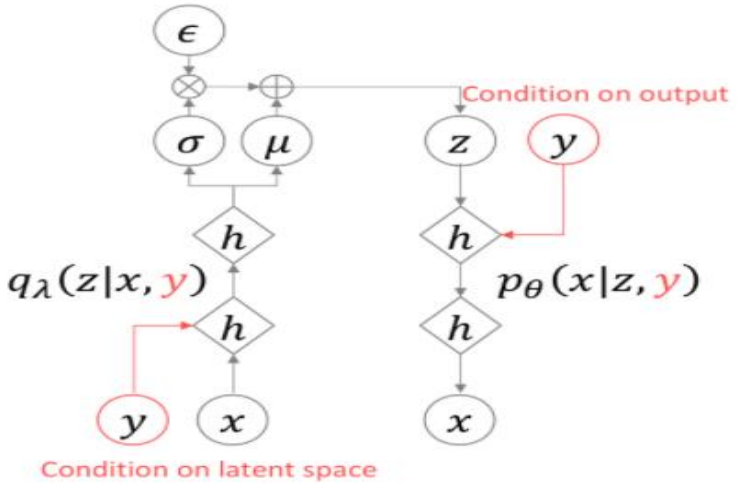
EVA_c의 장점

- **Supervised VAE**

① encoder, decoder에 정답 레이블 y가 추가된 형태

② z가 환자 고유의 특성 및 환자 별 condition에 따른 효과를 설명할 수 있음 (account for both effects arising from different medical conditions as well as those arising from individual differences)

But 많은 condition을 가진 환자를 고려할 수 없음



→ EVA_c is able to efficiently model such patients while learning **population wide medical condition representations** in addition to **patient specific** representations

3. Learning and Inference

Patient specific variables z_n, w_n, b_n 추정

- Posterior에 대한 근사 확률 분포 $q_\emptyset(z|x)$ 이용
- Inference network q에 sequence와 condition이 주어짐

$$q(z, w, b \mid \mathcal{D}, \{y_n\}_{n=1}^N) = \prod_{n=1}^N q(z_n, w_n, b_n \mid x_{n,1:T_n}, y_n)$$

$$= \prod_{n=1}^N \prod_{a \in \{z,w,b\}} q_{\phi_a}(a_n \mid x_{n,1:T_n}, y_n),$$

How to concatenate sequence and clinical condition vector? → Products of Expert

- 여러 간단한 확률 분포(experts)의 결과를 AND 연산으로 결합하여 복잡한 확률 분포를 모델링
- 개별 전문가가 저차원 공간에서 결정을 내릴 수 있게 하는 **Multi-modal learning**

  * e.g) 영상 데이터를 음향 전문가, 화질 전문가의 의견을 종합하여 분석

공통 target

$$q_{\phi_z}(z_n \mid x_{n,1:T_n}, y_n) = q(z_n \mid x_{n,1:T_n}) \cdot q(z_n \mid y_n)$$

설명 data

sequence experts × clinical condition experts

$$= N[z_n \mid LSTM_\mu(x_{n,1:T_n}), LSTM_\sigma(x_{n,1:T_n})] \times N[z_n \mid MLP_\mu(y_n), MLP_\sigma(y_n)]$$

Global variables $\mathbf{H}$, θ 추정

- **SG-MCMC** (Stochastic Gradient Markov Chain Monte Carlo)

① **Monte Carlo** : 랜덤 표본을 뽑아 함수의 값을 확률적으로 계산 → $q(z)$를 설정하는 것이 자유롭게 됨

$$q(z|x) = N\left(\mu_q(x), \Sigma_q(x)\right) \rightarrow x\text{가 달라지면 } q\text{의 분포도 달라짐}$$

$$z = \mu(x) + \sigma(x) \times \epsilon, \quad \epsilon \sim N(0,1) \rightarrow q\text{로부터 } z\text{를 직접 샘플링 하지 않음 (noise를 샘플링)}$$

② **Variational Inference with SGD** : q가 p를 잘 근사하도록 파라미터를 업데이트

③ **Markov Chain** : 특정 상태의 확률은 오직 과거 n개의 상태에 의존한다는 성질을 가진 확률 과정

4. Related Work

Generative Models

- **VAE**가 likelihood function에 대한 최적화를 통해 data의 분포를 학습, sampling은 분포에 의해 저절로 따라옴
- **GAN**은 likelihood free, 진짜 같은 sample을 확률적으로 생성하는 것이 목적

Conditional variants

- 미분 불가능한 점이 있는 **이산(discrete)** 데이터에 대해 GAN은 제대로 대처하지 못함
- VAE는 데이터의 확률 분포를 학습하기 때문에 **불연속적인 sequence를 생성**할 수 있음 (data에 대한 적절한 이산 확률 분포를 구체화할 수 있음)
- 특히 **Autoregressive** 분포와 결합되었을 때 효과적 (account for correlations exhibited by the data and prove convenient for specifying flexible densities over spatio-temporal sequences)

4. Related Work

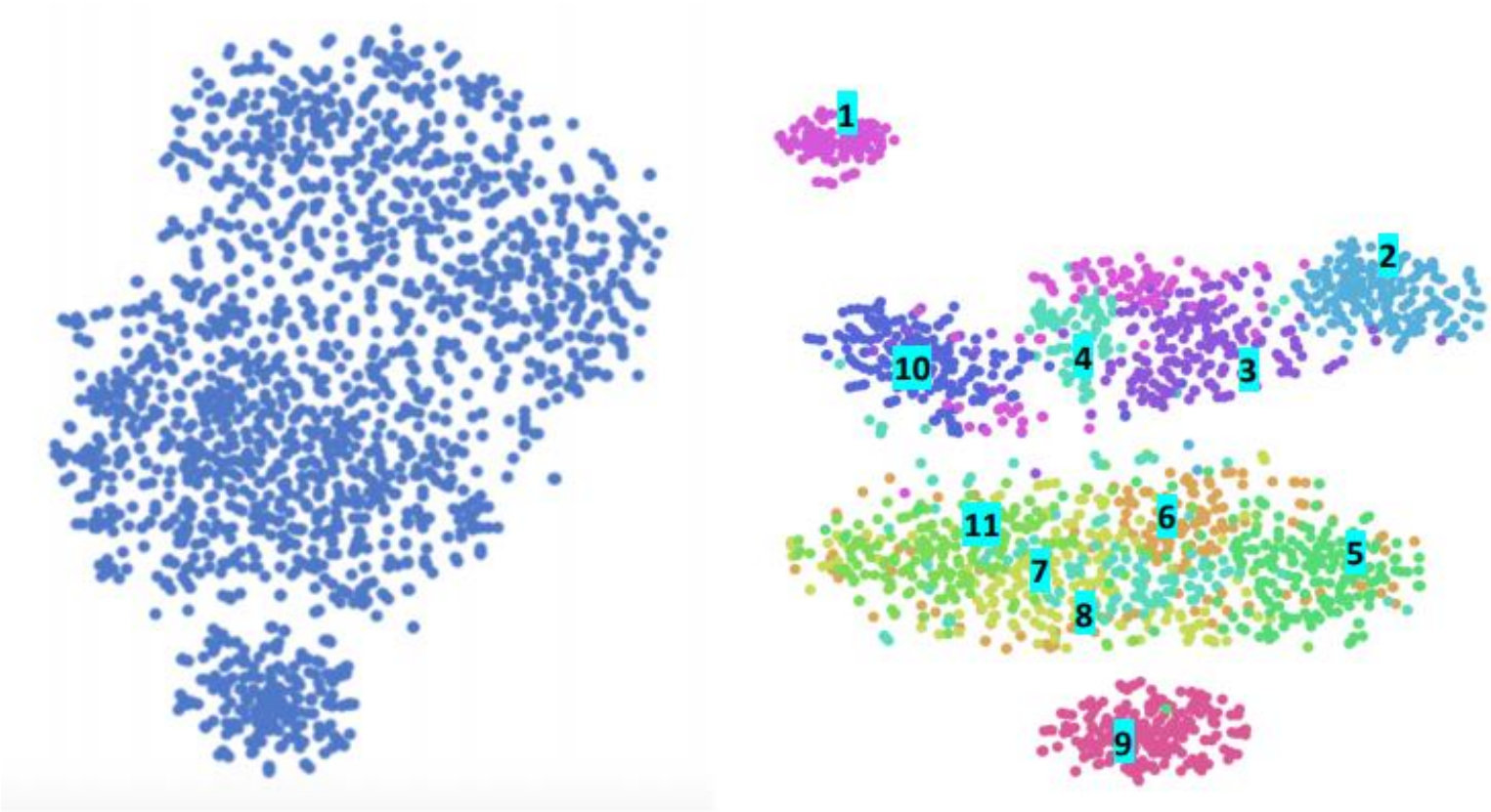
Conditional variants

- **Hierarchically factorized latent variables** by global parameter θ
- 환자 간의 개별적 차이에서 비롯되는 medical condition에 따른 variation을 잘 설명할 수 있음 (disentangling factors of variations stemming from conditions from those arising from individual differences among patients)
- 환자의 **meta data**를 활용해 **population wide latent variable**까지 고려

Synthetic EHR Generation

- Hand engineered rule에 의해 생성된 데이터보다 **일반화 가능성 높음, 생성이 용이**
- 현실 EHR sequence의 **불연속적인(temporal) 특징**을 반영하여 현실성을 높임

4. Related Work



- Latent space learned by EVA (left) and EVA_c (right)

→ EVA_c는 관심의 대상이 되는 medical condition에 특화된 잠재 벡터 학습

5. Experiments

Considerations

① EHR 통계량 : 생성된 EHR 데이터가 현실을 얼마나 잘 반영하는 지

② 생성된 EHR 데이터의 유용성

③ 사생활 보호 측면에 대한 평가

Methods for Comparison

① LSTM : 일종의 언어 생성 모델

② VAE-LSTM : VAE with LSTM decoder

③ VAE-Deconv : 위 모델의 LSTM을 Deconvolution network로 대체

④ EVA : 본 연구에서 제안된 모델

⑤ EVA_c : conditional variant of EVA (10개의 condition 고려)

Capturing EHR statistics

① **EHR 통계량**(predictive log likelihoods)을 기준으로 주어진 모델들을 평가

→ 본 연구에서 제안된 모델인 EVA, EVA_c가 높은 test 성능을 보임

② sequence 데이터를 생성, **bi-gram token의 발생에 대한 주변 확률** 계산

→ bi-gram에 속한 2개 단어들의 상관성 (얼마나 자연스러운 sequence를 생성하는 지 평가)

EVA: Generating Longitudinal Electronic Health Records Using Conditional Variational Autoencoders , ,

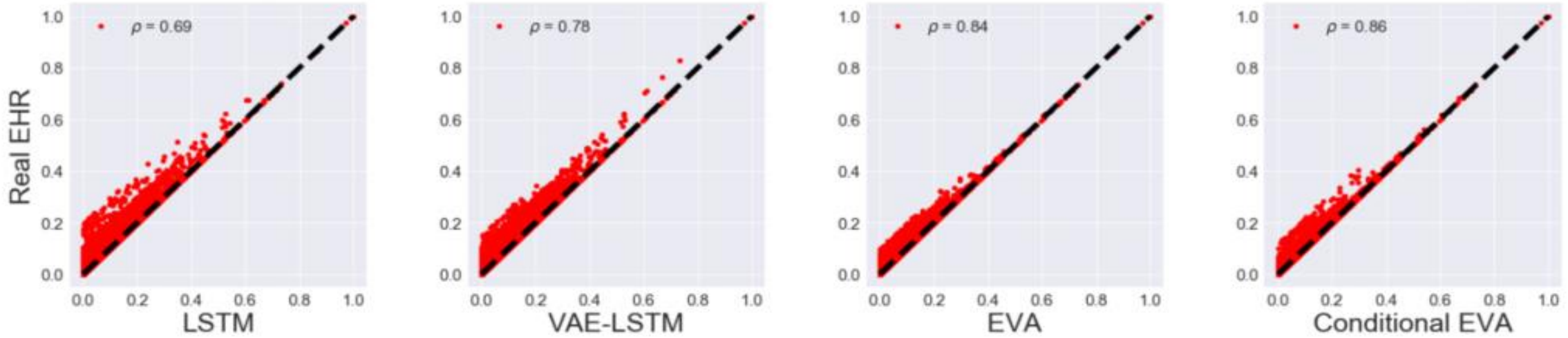


Figure 2: Comparison of bigram statitics of generated and real EHR codes. It confirmed that EVA generates data that capture better correlations in the EHR data.

5. Experiments

Usefulness of synthetic EHRs

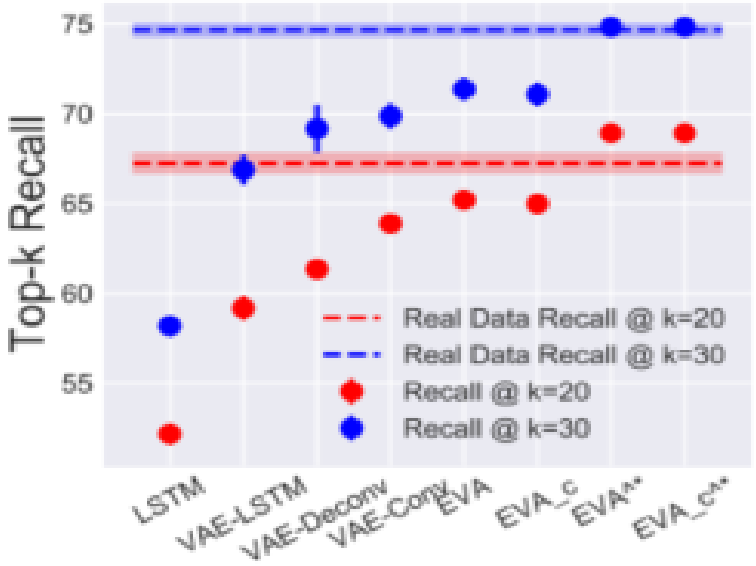
- 생성된 EHR sequence 데이터가 생존 분석에도 유용한지?

→ $\boldsymbol{predicting}\ \boldsymbol{P}(\boldsymbol{x}_{n,t} \mid \boldsymbol{x}_{n,t-1}, \boldsymbol{x}_{n,t-2}, \ldots, \boldsymbol{x}_{n,1})$: 환자의 history sequence가 주어졌을 때 미래 시점 예측

Accurate temporal prediction

- **Top k recall metric** : 관련 있는 상위 k개 예측에서의 TP / 총 TP의 개수 (본 연구에서는 20, 30 설정)

* '추천 시스템의 맥락에서 우리는 상위-K(top-K) 항목을 사용자에게 추천하는데 대체로 관심이 있으므로 **모든 항목이 아닌 상위 K개의 항목에 대한 정밀도와 재현율을 계산하는 것이 낫다.** 따라서 *K에서의 정밀도와 재현율* 이라는 표현에서 K는 상위-K 추천의 목적에 부합하기 위해 사용자가 정의할 수 있는 정수형 변수이다.'



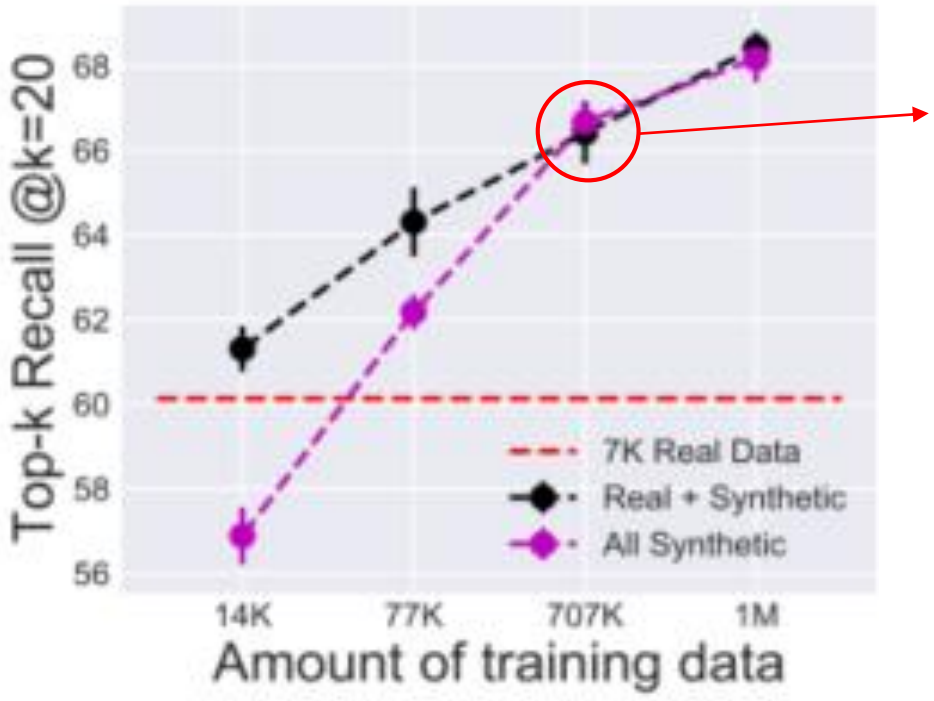
→ 소규모의 real 데이터를 사용하는 것보다 대용량의 synthesized 데이터를 활용하는 것이 보다 나은 성능을 보임

Beating Data via Data Augmentation

- 제한된 규모의 real EHR 데이터를 생성된 데이터로 **증식**시키는 것이 효과적인지?

① Synthesized 데이터만 사용했을 때는 real 데이터의 성능을 따라잡기 위해서는 대용량의 데이터가 필요

② **기존의 real 데이터에 synthesized 데이터를 결합시켰을 때는 항상 real 데이터의 성능을 초과함**



전체 데이터 셋 중 synthesized 데이터가 차지하는 비중이 증가함에 따라, real 데이터의 dominance가 낮아짐

Benefits of Conditional generation

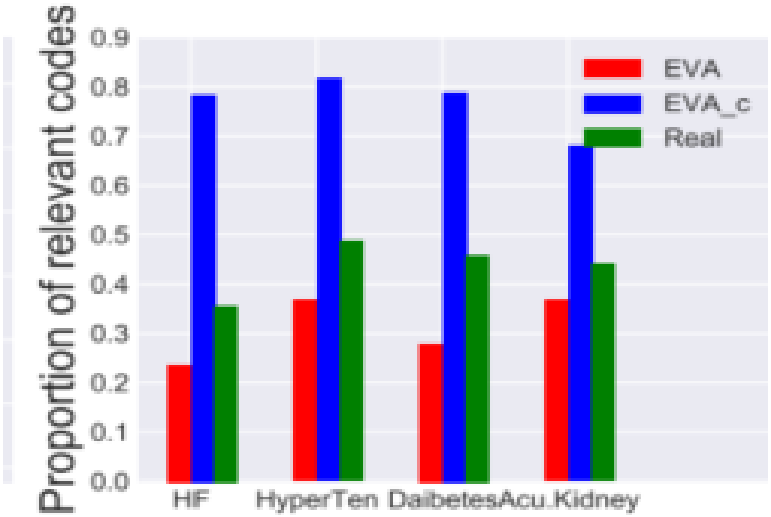
- Conditional model은 **특정 condition**에 specific한 EHR sequence를 생성할 수 있음 - EVA_c

  e.g) 과거의 history sequence가 주어졌을 때 심장병 발병을 예측할 수 있을까?

  → 심잠병을 앓고 있는 환자들, 그렇지 않은 환자들에 대한 데이터가 필요

- **EVA는 일반적인 sequence 데이터밖에 생성하지 못함** (Generation of data with **cases and controls** cannot be achieved using EVA) , noise sampling

  ↔ EVA_c는 심장병에 **specific**한 sequence 데이터를 생성할 수 있음



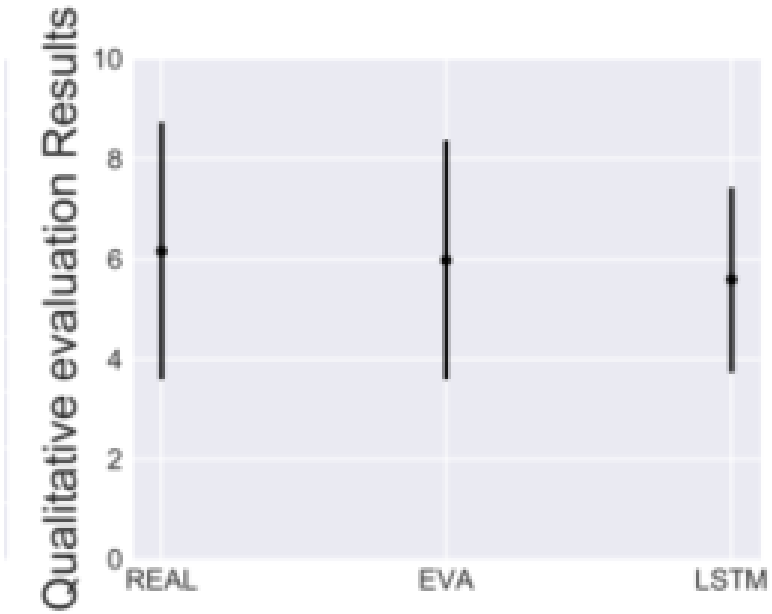
→ **특정 condition에 specific한 real 데이터가 있을 때,**
EVA는 그러한 특성을 제대로 반영하지 못함

5. Experiments

Clinical User Study

- 생성된 EHR sequence 데이터에 대한 질적 평가 → 임상 전문가들이 synthesized 데이터의 현실성 평가 (0~10)
- EVA는 충분히 현실적인 synthesized 데이터를 제공함

  (평균 6점 : It is mainly due to the details associated with EHR structured data are not available)



→ 평가 점수에 대한 plot

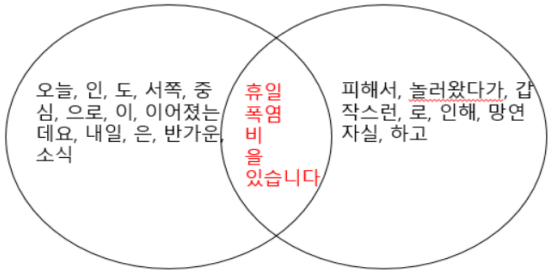
5. Experiments

Diversity within visits

- 개별 환자의 sequence 데이터의 points(병원 방문)에 대한 다양성 평가 (diversity of visits within a single patient in our data)

$$x_{n,1:T_n} = \{x_{n,1}, x_{n,2}, \ldots, x_{n,T_n}\}$$

- 방문간의 **자카드 유사도** 계산 : avg $[similarity(x_{n,1}, x_{n,2}),\ similarity(x_{n,2}, x_{n,3}), \ldots similarity(x_{n,t-1}, x_{n,t})]$



→ 두 집합의 교집합 / 두 집합의 합집합

Table 1: Average Jaccard Similarity Index

Model	Jaccard Coefficient
LSTM	0.3874
VAE-LSTM	0.3156
VAE-Deconv	0.2631
EVA	0.2167
EVA_c	0.2235
Real EHR	0.1835

→ **EVA, EVA_c가 현실의 데이터와 유사한 다양성이 높은 sequence 데이터를 생성**

5. Experiments

Diversity within visits – weight uncertainty in θ

$$p(\mathcal{D}, \mathcal{Z}, \theta, \mathbf{H} \mid \eta) = p(\theta)p(\mathbf{H}) \prod_{n=1}^{N} p(w_n)p(b_n \mid \gamma)$$
$$p(z_n \mid \mathbf{H}, \pi_n, b_n, \tau)p(x_{n,1:\mathrm{T}_n} \mid f_\theta(z_n)),$$

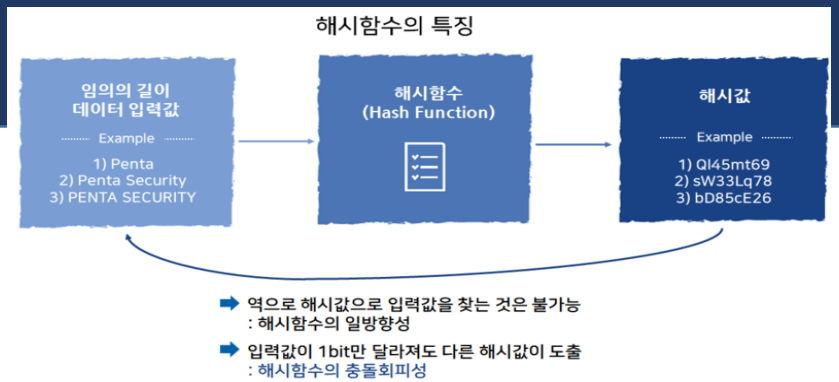
- Global parameter에 대한 point estimation보다 **확률 분포를 추정**하는 것이 다양하고 현실적인 sequence 데이터를 생성하는데 유리
- The point estimate variants produce sequences with unnaturally many **repeated tokens in a sequence**

→ 평가 지표 : # unique tokens / # total tokens

Point estimate with	Score
EVA	0.2786 ± 0.08
Bayesian	0.3214 ± 0.07
Real EHR	0.3845 ± 0.08

→ EVA를 활용해 sequence 데이터를 생성할 때 θ에 대한 점 추정치를 활용하면 real sequence가 가진 다양성을 제대로 표현하지 못함

5. Experiments



Privacy Risk Evaluation

- 생성된 데이터의 개인정보 보호 측면을 평가
- EVA를 통한 sequence 데이터 생성이 생성된 데이터와 원본(훈련) 데이터 사이의 **1-1 mapping을 불가능하게 해야 함**

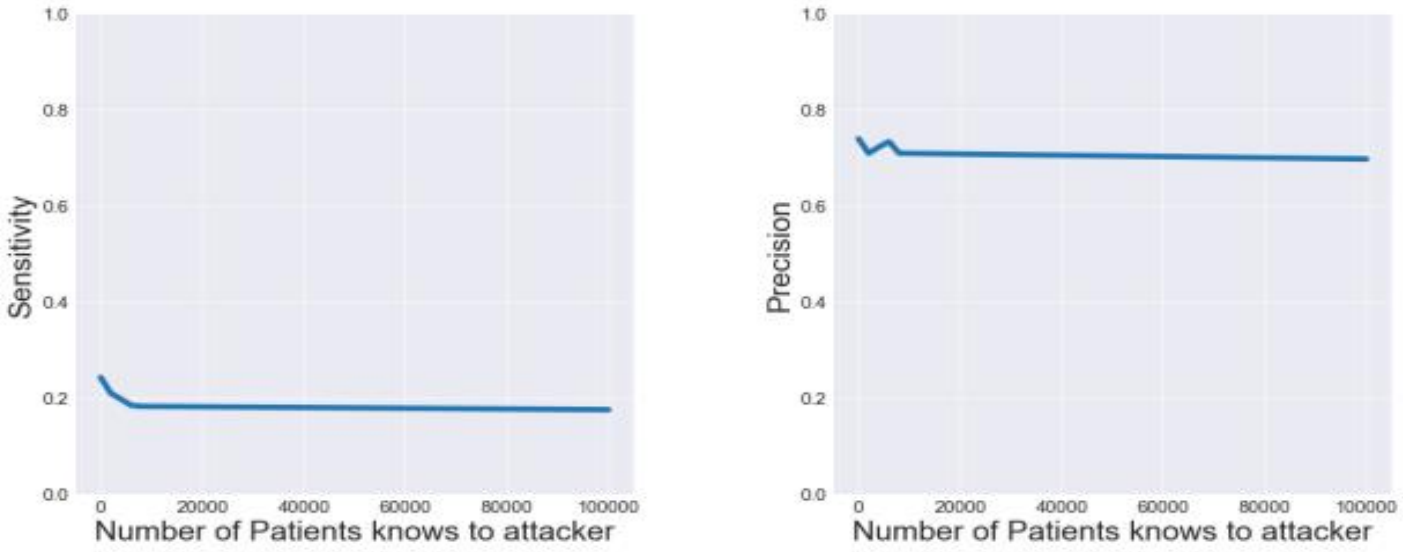
Privacy Risk Evaluation – Presence disclosure

- EVA가 환자 a의 기록을 포함한 데이터 셋을 통해 훈련되었다는 것을 알 수 있으면 재식별화 발생
- 해커가 개인 a를 특정하려는 시도의 성공에 대한 평가 지표로 **민감도, 정확도** 이용
  - 80% 민감도 : 해커가 이미 알고 있는 환자들 중 80%가 실제로 EVA를 학습시키는데 사용되었음
  - 80% 정확도 : 해커가 특정 환자들이 실제로 EVA를 학습시키는데 사용되었다고 주장할 때, 그 중 80%가 실제로 학습에 사용되었음

① a가 학습 데이터에 존재한다는 것에 대한 사전 확률=0.8 (가정)

② **해커가 민감도와 정확도를 0.8이상으로 개선할 수 있는 지 판별** (information gain 여부를 판별)

5. Experiments



(a) Sensitivity (b) Precision

Figure 5: Sensitivity and precision vs. number of patients known to the attacker.

① 해커가 알고 있는 환자들 중 20%만이 실제로 EVA를 훈련시키는데 사용되었음

② 정확도가 0.8을 초과하지 못함 (< 사전확률=0.8)

③ 해커가 몇 명의 환자들에 대한 정보를 소유하고 있는 지와 관계없이 생성 데이터는 해커의 공격으로 부부터 안전함

6. Conclusions

의의

1. 진짜 같은 가짜 sequence 데이터의 생성이 가능
2. 헬스케어 영역에 대용량의 비식별화된 데이터를 제공하고, 이를 바탕으로 AI 기술을 활용할 수 있게 함
3. 실제 데이터를 사용하는 것 이상의 모델 예측 성능을 보장

개선점

- Sequence point간의 시차를 고려 (time gap between clinical visits) → 임상 이외의 요인들에 영향을 받음
- 추가적인 meta data를 고려하면 inter arrival time of clinical visits를 모델링 할 수 있을 지도?

참고 문헌

1. VAE : https://ratsgo.github.io/generative%20model/2018/01/27/VAE/
2. Variational Inference : https://ratsgo.github.io/generative%20model/2017/12/19/vi/
3. Conditional VAE : https://ratsgo.github.io/generative%20model/2018/01/28/VAEs/