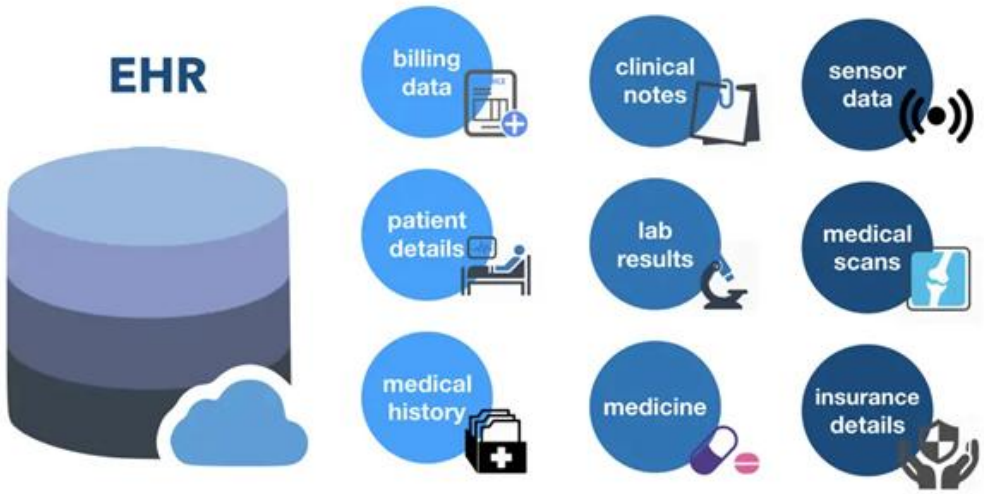
电子健康记录中的数据规律

潘达岩 @ BIGSCity

目录

H. Ren, J. Wang, X. W. Zhao, and N. Wu, " RAPT: Pre-training of Time-Aware Transformer for Learning Robust Healthcare Representation," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining (KDD'21),* 2021, pp. 3503-3511.

什么是EHR数据



- **电子健康纪录**，简称EHR (electronic health record)，是电子化的个人健康纪录（病历、心电图、医疗影像等）

$$\boldsymbol{e} = (e_1, \ldots, e_m, \ldots, e_{Me})$$

什么是EHR数据

肝功全项
生化组 2023-01-19

急诊尿常规十二项
住院急诊组 2023-01-19

肾功+血钙+钾钠氯+葡萄糖
住院急诊组 2023-01-23

急诊C-反应蛋白(CRP)+血常规
住院急诊组 2023-01-23

肝功全项
生化组 2023-01-24

肝功全项+肾功全项+血钙磷+钾钠氯+血糖
生化组 2023-01-27

CRP
免疫组 2023-01-27

血常规五分类
临检组 2023-01-27

血常规五分类
临检组 2023-01-30

CRP
免疫组 2023-01-30

普外二科
CRP

送检医生 送检时间 2023-01-27 09:41:29

检查项目	结果	参考范围
C反应蛋白	37.30mg/L ↑	0.00 ~ 8.00

报告医生 报告时间 2023-01-27 09:59:12

数据缺失

普外二科
CRP

送检医生 送检时间 2023-01-30 08:06:54

检查项目	结果	参考范围
C反应蛋白	19.70mg/L ↑	0.00 ~ 8.00

报告医生 报告时间 2023-01-30 09:36:38

普外二科
肝功全项

送检医生 送检时间 2023-01-19 08:13:50

检查项目	结果	参考范围
α1球蛋白	8.5% ↑	2.9 ~ 4.9
α2球蛋白	15.5% ↑	7.1 ~ 11.8
β1球蛋白	6.1%	4.1 ~ 7.2
β2球蛋白	7.0% ↑	3.2 ~ 6.5
γ球蛋白	15.3%	

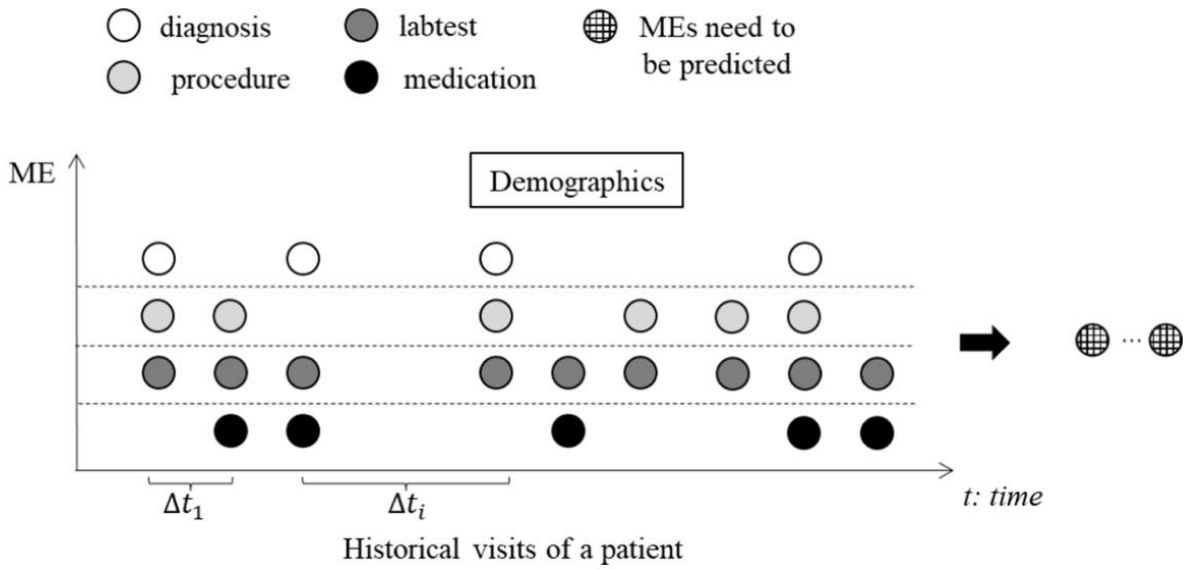
难以理解

普外二科
肝功全项+肾功全项+血钙磷+钾钠氯+血糖

送检医生 送检时间 2023-01-30 10:16:01

检查项目	结果	参考范围
α1球蛋白	6.3% ↑	2.9 ~ 4.9
α2球蛋白	14.3% ↑	7.1 ~ 11.8
β1球蛋白	6.3%	4.1 ~ 7.2
β2球蛋白	6.3%	3.2 ~ 6.5
γ球蛋白	16.9%	11.1 ~ 18.8

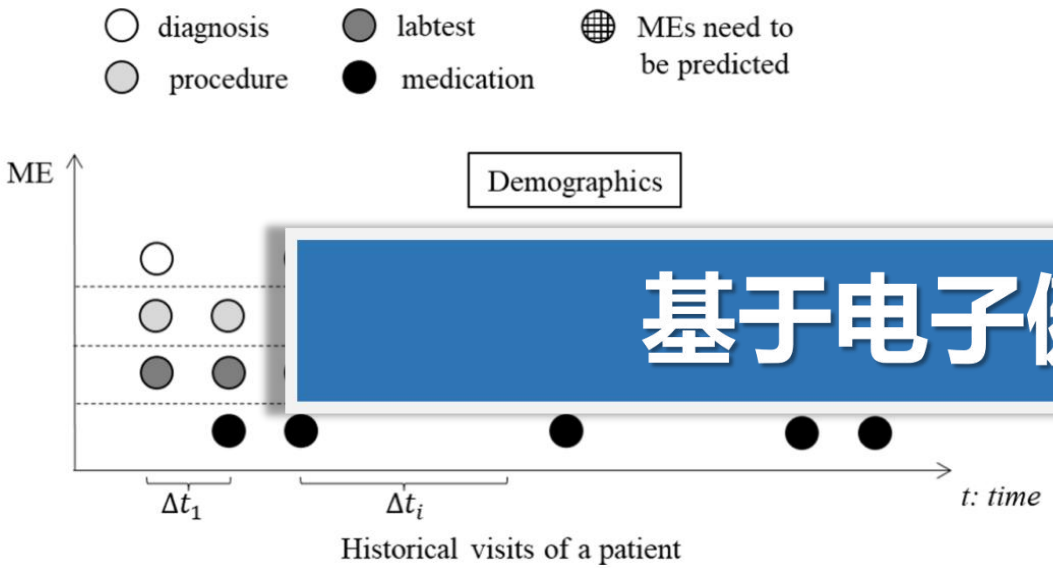
EHR数据的特点与挑战



利用EHR数据的挑战	
理解困难	从EHR数据中不能直接理解数据规律
标签稀疏	检验的结果一般不会体现在报告上
缺失严重	患者不会每次进行所有检查
风险极高	误诊风险与隐私风险极大地限制了预测工作

S. Liu, X. Wang, Y. Xiang, H. Xu, H. Wang, and B. Tang, "Multi-channel fusion LSTM for medical event prediction using EHRs," *Journal of Biomedical Informatics*, vol. 127, p. 104011, 2022.

EHR数据的特点与挑战



利用EHR数据的挑战

理解困难 从EHR数据中不能直接理解数据规律

缺失严重 患者不会每次进行所有检查

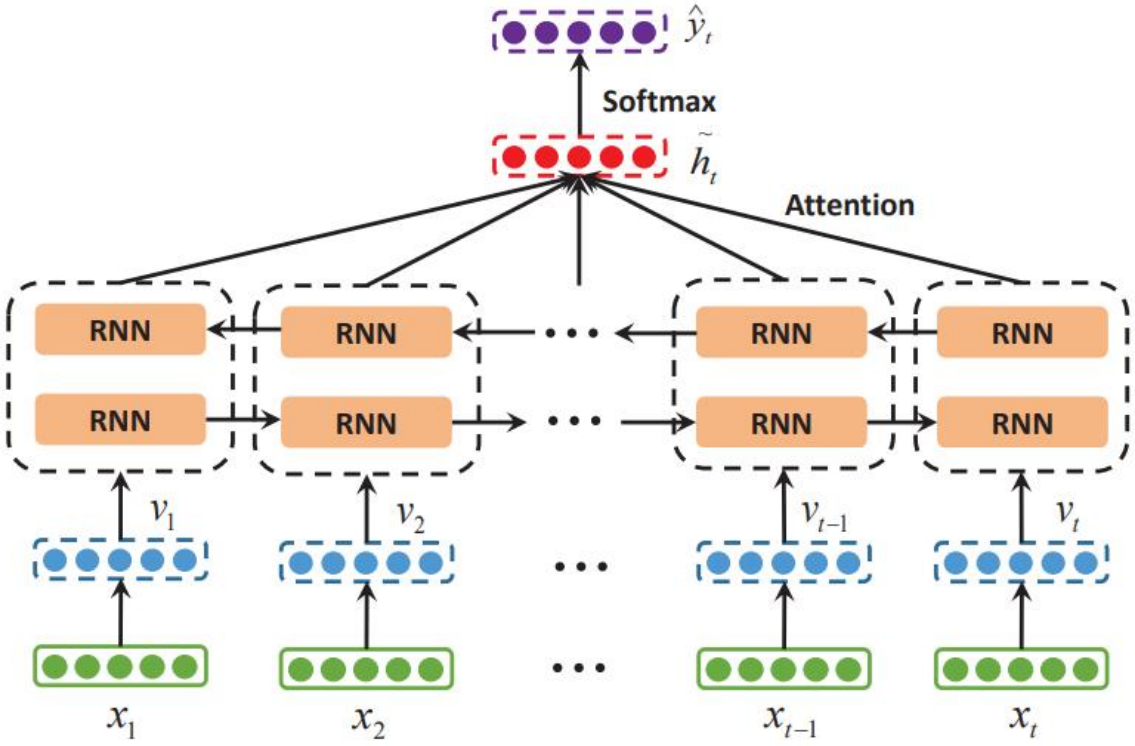
风险极高 误诊风险与隐私风险极大地限制了预测工作

基于电子健康记录的数据挖掘

基于电子健康记录的数据挖掘

利用EHR数据的挑战		
理解困难	从EHR数据中不能直接理解数据规律	序列数据预测
标签稀疏	检验的结果一般不会体现在报告上	自监督学习
缺失严重	患者不会每次进行所有检查	异构数据&数据生成
风险极高	误诊风险与隐私风险极大地限制了预测工作	隐私保护&可解释性

电子健康记录中的序列数据预测方法



序列数据预测

双向LSTM直接预测

F. Ma, R. Chitta, J. Zhou, Q. You, T. Sun, and J. Gao, "Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks," in *Proceedings of the 23th ACM SIGKDD Conference on Knowledge Discovery & Data Mining (KDD'17)*. ACM, 2017, pp. 1903–1911.

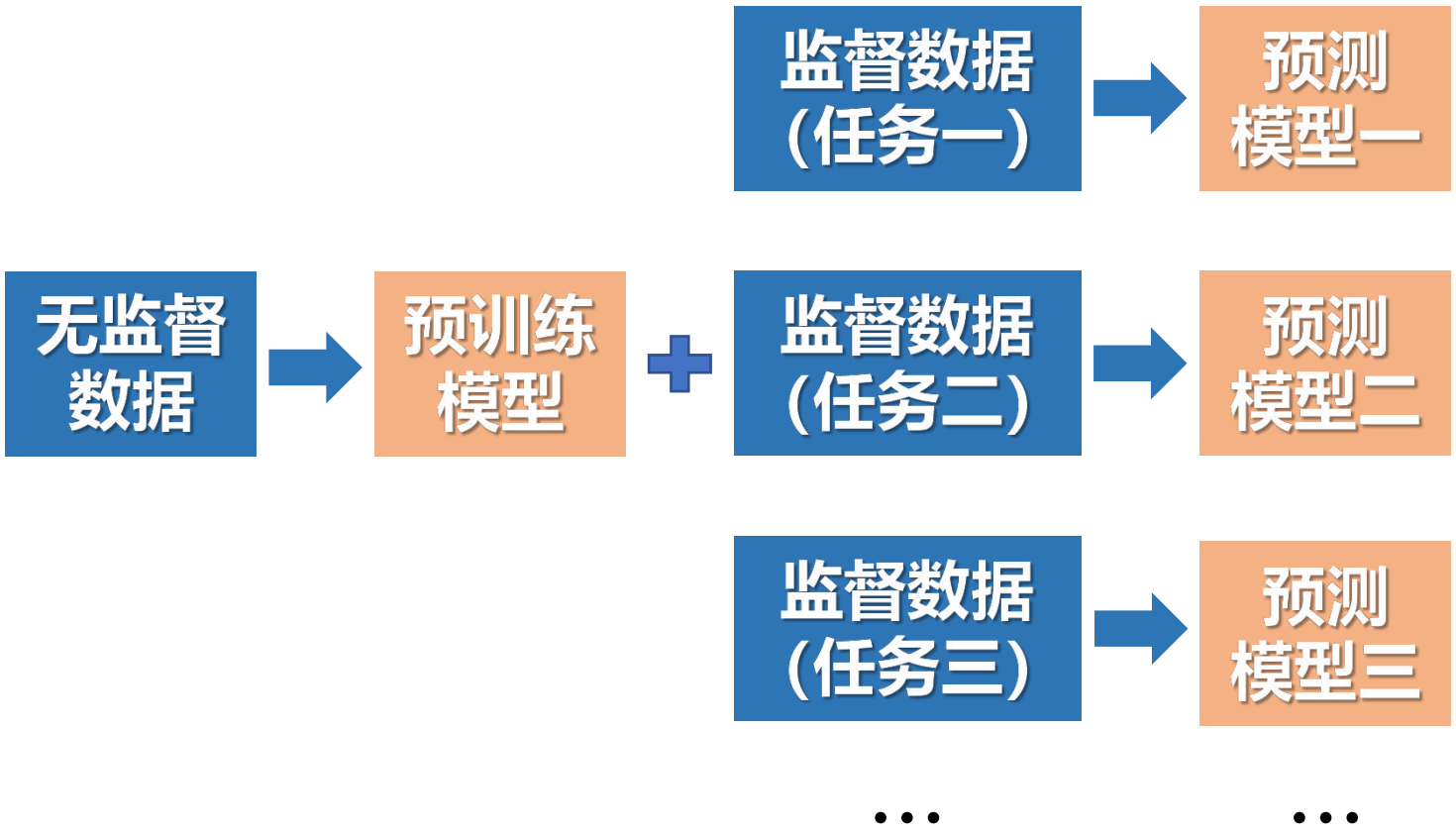
基于电子健康记录的数据挖掘

利用EHR数据的挑战		
理解困难	从EHR数据中不能直接理解数据规律	序列数据预测
标签稀疏	检验的结果一般不会体现在报告上	自监督学习
缺失严重	患者不会每次进行所有检查	异构数据&数据生成
风险极高	误诊风险与隐私风险极大地限制了预测工作	隐私保护&可解释性

基于预训练的电子健康记录表征学习

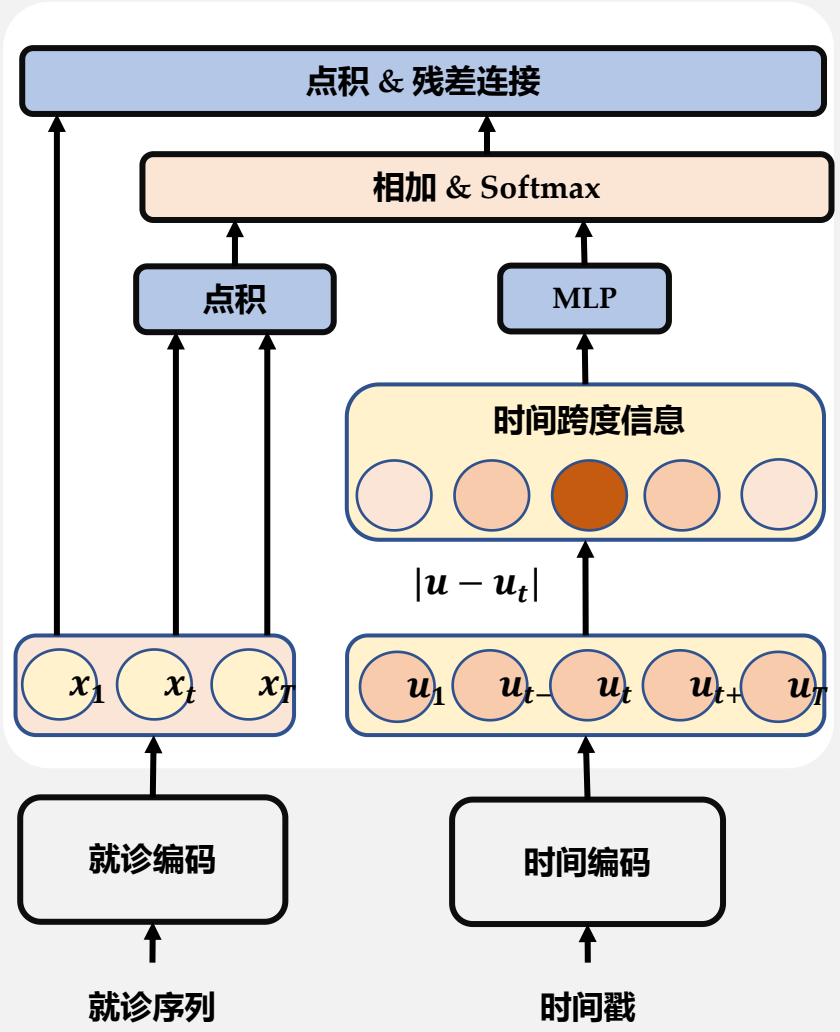
H. Ren, J. Wang, X. W. Zhao, and N. Wu, " RAPT: Pre-training of Time-Aware Transformer for Learning Robust Healthcare Representation," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining (KDD'21),* 2021, pp. 3503-3511.

背景



- 设计一个适合建模电子健康记录和预训练的网络架构。
- 设计合适的预训练任务，模型可以通过这些任务学习表征。

模型结构



- 标准自注意力机制：

$$a_{i,j} = \frac{q_i^T k_j}{\sqrt{h}} = \frac{x_i^T W_q^T W_k x_j + x_i^T W_q^T W_k u_j + u_i^T W_q^T W_k x_j}{\sqrt{h}}.$$

- 时间感知自注意力机制：

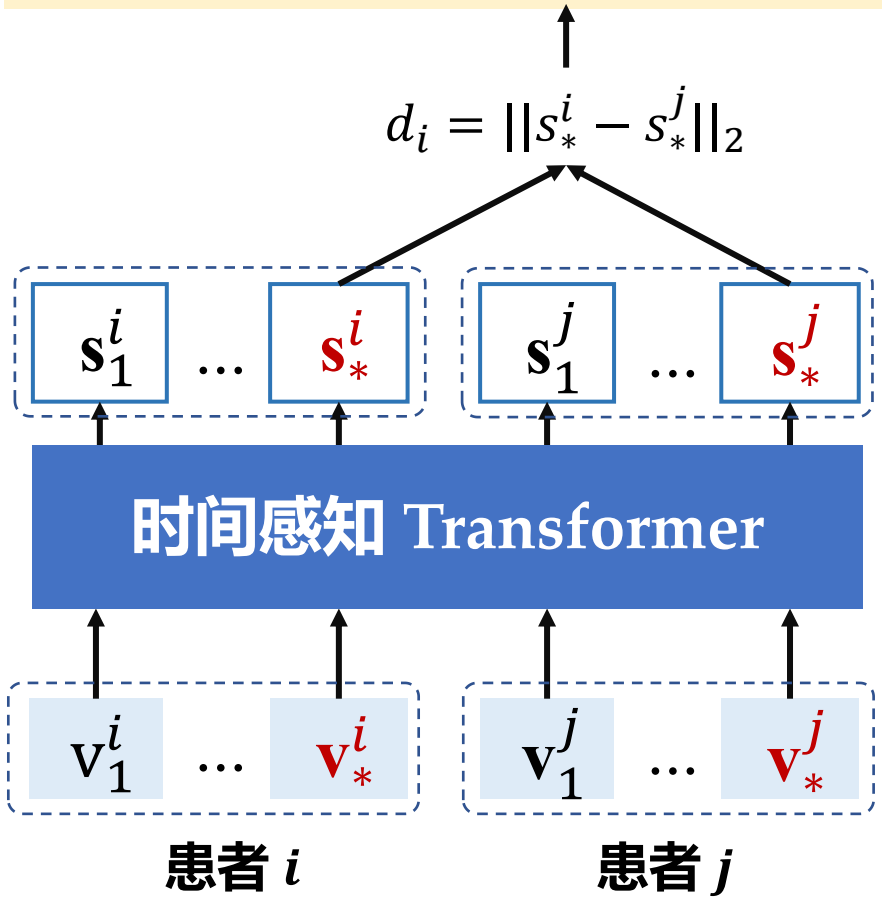
$$a_{i,j}^{(T)} = \frac{q_i^T k_j}{\sqrt{h}} = \frac{x_i^T W_q^T W_k x_j + W_\tau |u_i - u_j|}{\sqrt{h}},$$

- 其中，W_τ是可学习的参数，$|\cdot|$表示元素级别的绝对值。在此，试图通过$W_\tau |u_i - u_j|$建模两次就诊的时间间隔。

预训练任务

任务一
相似病例预测

$$\mathcal{L}_s = \left(z_i d_i^2 + (1 - z_i)\max\left(m - d_i^2, 0\right)^2\right)$$

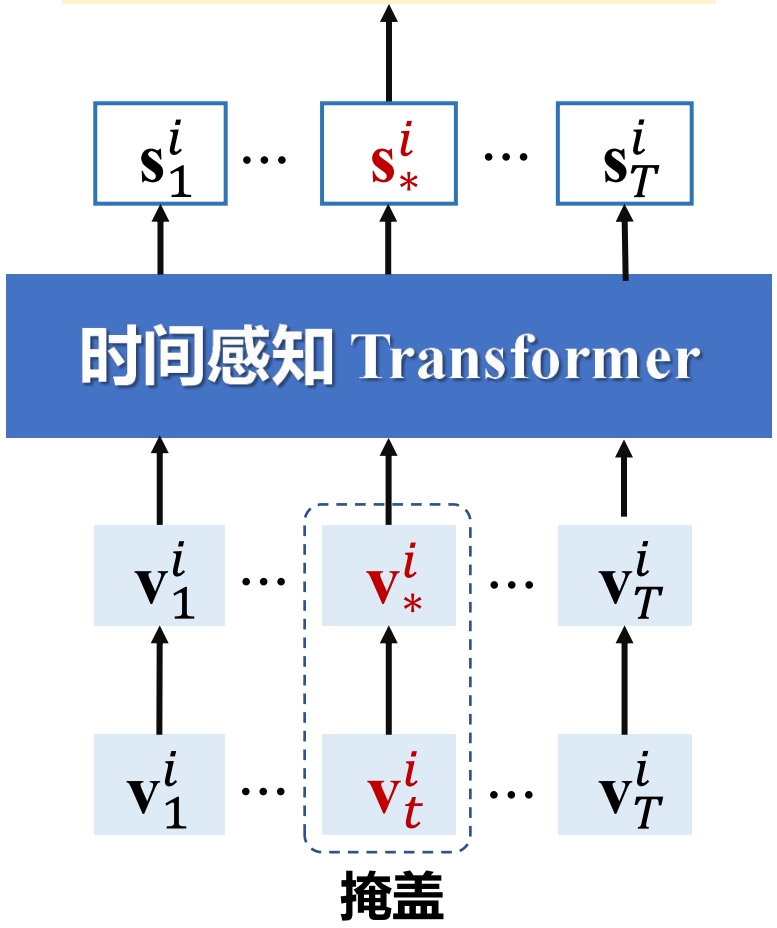


- 统计所以电子健康记录的欧式距离
- 取最大的15%作为负样本
- 取最小的15%作为正样本
- 对比学习优化模型

预训练任务

任务二
掩盖记录预测

$$\mathcal{L}_m = ||\mathbf{c}_t^i - \mathrm{MLP}(\mathbf{s}_*^i)||_2^2$$

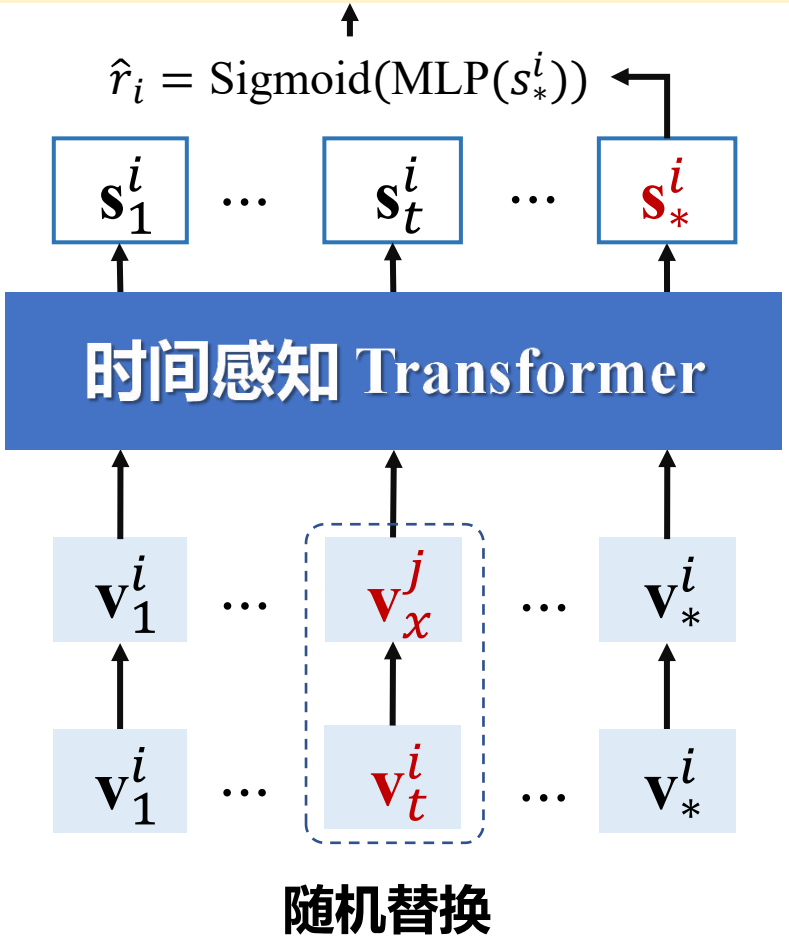


- 随机掩盖30%的就诊
- 将掩盖之后的就诊序列输入到模型中
- 通过对应的就诊表征来预测（利用上下文）被掩盖的就诊
- 均方误差优化模型

预训练任务

任务三
合理记录趋势检查

$$\mathcal{L}_r = \mathrm{r}_i \log(\hat{\mathrm{r}}_i) + (1 - \mathrm{r}_i)\log(1 - \hat{\mathrm{r}}_i)$$



随机替换

- 负样本：随机将50%的就诊替换为其他就诊序列的就诊，改变序列的合理性
- 正样本：不做任何改变
- 预测序列是否合理
- 交叉熵优化模型

训练流程

- 使用三个预训练任务优化模型
- 丢弃预训练任务使用的分类头
- 随机初始化下游任务的分类头
- 使用下游任务数据微调模型

算法 1 RAPT 训练算法

输入： 电子健康记录数据集 $\mathbb{D}$

输出： 模型参数 θ_I，θ_T 以及 θ_F

随机初始化参数 θ_I，θ_T 以及 θ_P
while 预训练未收敛 **do**
    通过公式 (3.13)计算预训练损失
    由公式 (3.13)执行梯度下降算法，更新 θ_I，θ_T 以及 θ_P
end while
丢弃 θ_P
随机初始化参数 θ_F
while 微调未收敛 **do**
    通过公式 (3.15)或者公式 (3.17)计算微调损失
    由公式 (3.15)或者公式 (3.17) 执行梯度下降算法，更新 θ_I，θ_T 以及 θ_F
end while
输出： θ_I，θ_T 以及 θ_F

实验数据

数据集	预训练	糖尿病预测	高血压预测	妊娠结局预测	危险期预测
样例数量	63,001	20,160	5,744	8,514	1,556
就诊次数	427,369	137,873	38,600	57,081	19,434
平均就诊序列长度	6.78	6.84	6.72	6.70	12.49
平均首次就诊孕周	13.82	14.46	14.51	14.50	14.63
平均末次就诊孕周	28.18	28.23	28.21	28.20	36.96

- **妊娠糖尿病预测** 该任务旨在预测患妊娠糖尿病的概率。
- **妊娠高血压预测** 该任务旨在预测患妊娠高血压的概率。
- **妊娠结局预测** 该任务旨在预测最后一周的身体状况。
- **危险期预测** 该任务旨在诊断本周被诊断为妊娠高血压的概率。

- **数据集划分**
  - **训练集/验证集/测试集**：0.7/0.1/0.2
  - **五折交叉验证**
- **测试指标**
  - **分类任务**：F1、AUC-ROC、AUC-PR
  - **回归任务**：RMSE、R2、EV
- **对比方法**
  - **通用序列建模模型**：LSTM、Transformer
  - **电子健康记录建模模型**：RETAIN、T-LSTM、Dipole、HiTANet

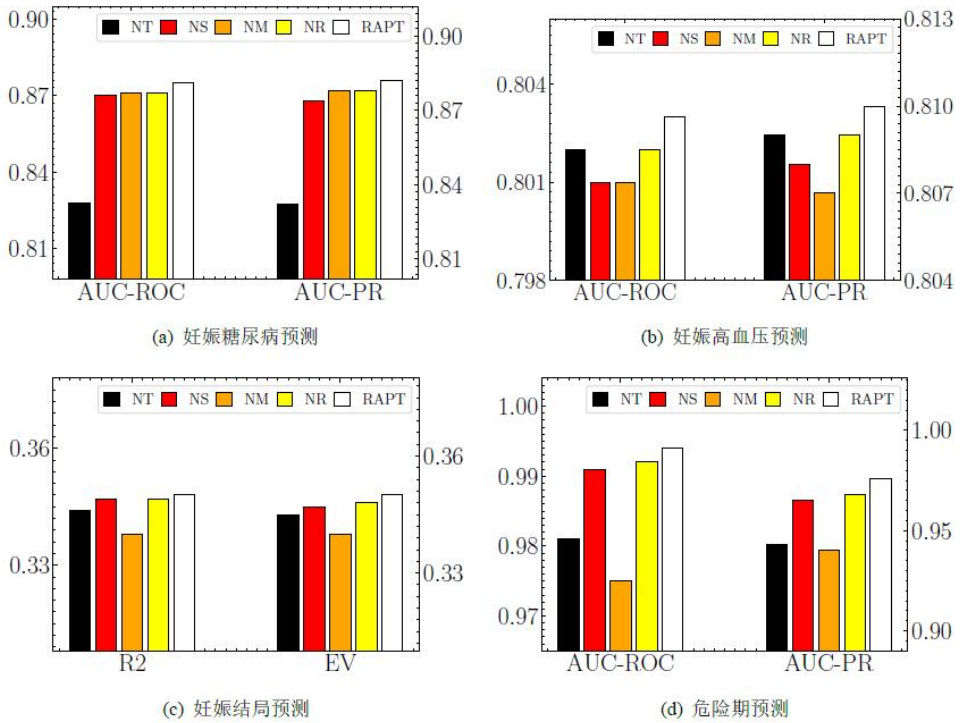
实验结果

任务		妊娠糖尿病预测		
评价指标		F1 ↑	AUC-PR ↑	AUC-ROC ↑
方法	LSTM	0.680 ± 0.019	0.719 ± 0.008	0.727 ± 0.007
	Transformer	0.725 ± 0.022	0.794 ± 0.040	0.797 ± 0.030
	RETAIN	0.663 ± 0.043	0.732 ± 0.012	0.751 ± 0.012
	TLSTM	0.669 ± 0.033	0.768 ± 0.008	0.781 ± 0.008
	Dipole	0.675 ± 0.041	0.739 ± 0.008	0.756 ± 0.006
	HiTANet	0.756 ± 0.013	0.823 ± 0.028	0.826 ± 0.020
	$RAPT_{NP}$	0.770 ± 0.003	0.843 ± 0.012	0.847 ± 0.006
	RAPT	**0.802 ± 0.007***†§	**0.882 ± 0.006***†§	**0.875 ± 0.006***†§
任务		妊娠高血压预测		
评价指标		F1 ↑	AUC-PR ↑	AUC-ROC ↑
方法	LSTM	0.711 ± 0.026	0.804 ± 0.016	0.797 ± 0.012
	Transformer	0.711 ± 0.021	0.801 ± 0.018	0.796 ± 0.015
	RETAIN	0.708 ± 0.018	0.800 ± 0.021	0.796 ± 0.015
	TLSTM	0.701 ± 0.029	0.807 ± 0.019	0.800 ± 0.016
	Dipole	0.714 ± 0.025	0.805 ± 0.017	0.799 ± 0.013
	HiTANet	0.709 ± 0.024	0.796 ± 0.017	0.793 ± 0.012
	$RAPT_{NP}$	0.702 ± 0.025	0.806 ± 0.018	0.799 ± 0.015
	RAPT	**0.734 ± 0.011***†§	**0.810 ± 0.014**†§	**0.803 ± 0.009***†

任务		妊娠结局预测		
评价指标		RMSE ↓	R2 ↑	EV ↑
方法	LSTM	10.274 ± 0.217	-0.002 ± 0.000	0.000 ± 0.000
	Transformer	7.984 ± 0.233	0.46 ± 0.014	0.48 ± 0.014
	RETAIN	8.489 ± 0.293	0.276 ± 0.016	0.277 ± 0.016
	TLSTM	10.269 ± 0.217	-0.001 ± 0.001	-0.001 ± 0.000
	Dipole	8.352 ± 0.231	0.296 ± 0.010	0.297 ± 0.010
	HiTANet	7.968 ± 0.225	0.48 ± 0.008	0.340 ± 0.008
	$RAPT_{NP}$	7.924 ± 0.276	0.340 ± 0.014	0.341 ± 0.015
	RAPT	**7.885 ± 0.230***†§	**0.348 ± 0.010***†§	**0.350 ± 0.010***†§
任务		危险期预测		
评价指标		F1 ↑	AUC-PR ↑	AUC-ROC ↑
方法	LSTM	0.742 ± 0.022	0.871 ± 0.009	0.957 ± 0.004
	Transformer	0.679 ± 0.108	0.851 ± 0.008	0.947 ± 0.004
	RETAIN	0.554 ± 0.064	0.751 ± 0.042	0.923 ± 0.014
	TLSTM	0.754 ± 0.030	0.877 ± 0.008	0.959 ± 0.004
	Dipole	0.782 ± 0.010	0.893 ± 0.005	0.965 ± 0.004
	HiTANet	0.771 ± 0.050	0.881 ± 0.037	0.962 ± 0.014
	$RAPT_{NP}$	0.784 ± 0.115	0.903 ± 0.040	0.967 ± 0.014
	RAPT	**0.948 ± 0.005***†§	**0.976 ± 0.010***†§	**0.994 ± 0.003***†§

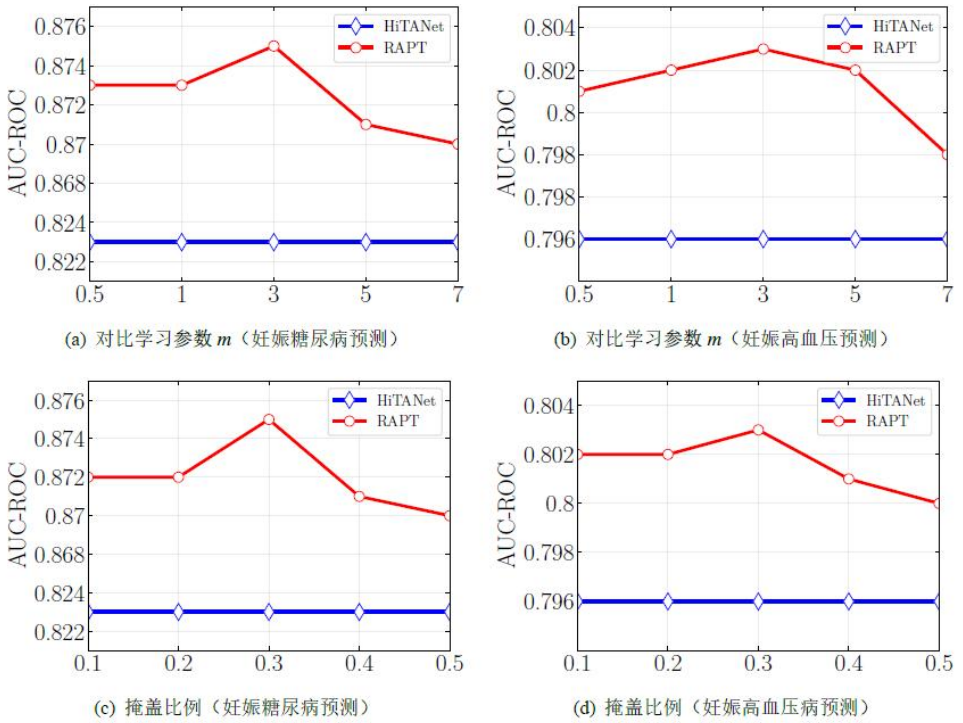
- 所提出的时间感知Transformer优于标准的Transformer
- 预训练之后的模型优于完全监督训练的模型

实验结果



(a) 妊娠糖尿病预测　(b) 妊娠高血压预测

(c) 妊娠结局预测　(d) 危险期预测

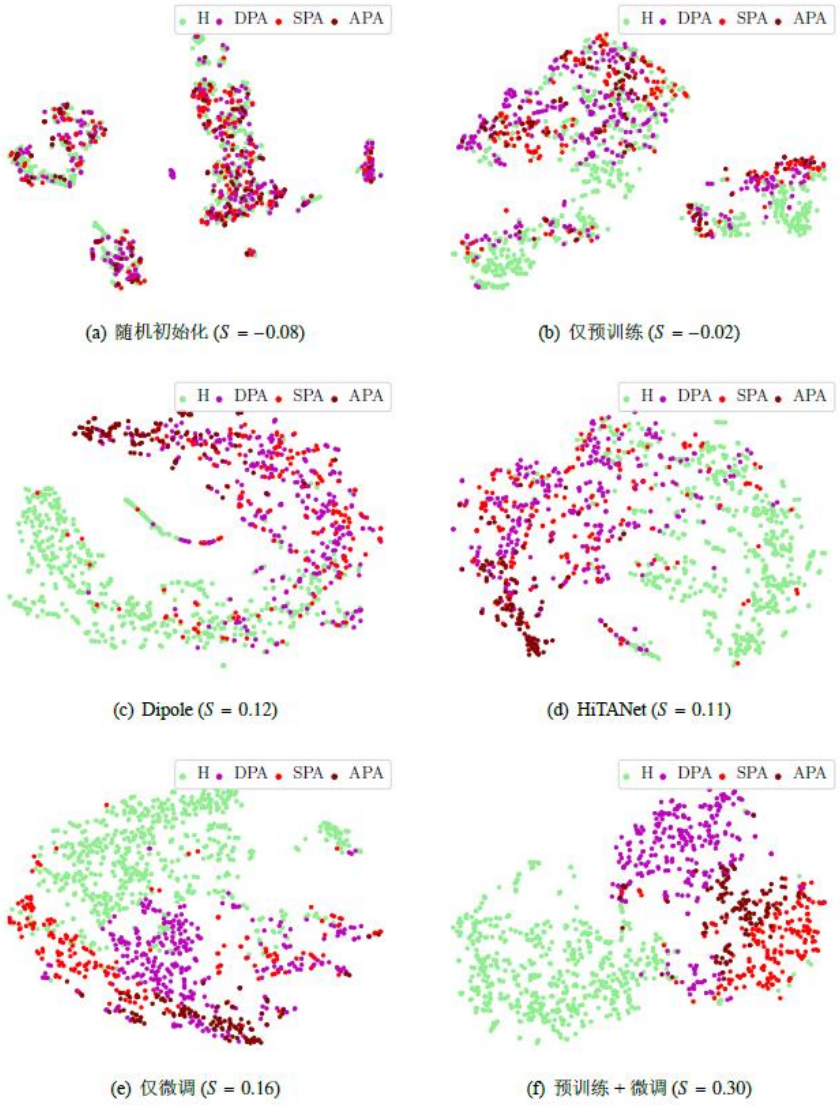


(a) 对比学习参数 m（妊娠糖尿病预测）　(b) 对比学习参数 m（妊娠高血压预测）

(c) 掩盖比例（妊娠糖尿病预测）　(d) 掩盖比例（妊娠高血压病预测）

- 消融实验：证明了所提出的预训练架构和三个预训练任务的有效性。
- 敏感性分析实验：证明了所提出方法的鲁棒性。

实验结果



(a) 随机初始化 (S = −0.08)

(b) 仅预训练 (S = −0.02)

(c) Dipole (S = 0.12)

(d) HiTANet (S = 0.11)

(e) 仅微调 (S = 0.16)

(f) 预训练 + 微调 (S = 0.30)

- 随机初始化：属于不同类别的点是完全杂乱无章的。
- 仅经过预训练时间感知Transformer：健康样例和高血压样例略有分离，虽然结果仍不尽如人意，但考虑到没有任何监督信号，结果是可以接受的。
- Dipole、HiTANet、仅经过微调的时间感知Transformer：可以区分高血压和健康样例
- 经过预训练和微调的时间感知Transformer：可以进一步区分不同症状的高血压样例。
- 对比不同模型的Silhouette分数也说明了该模型效果最好。
- 综合上述分析可以发现，经过预训练之后，模型可以学习到更加有效鲁棒的表征。

目录

基于电子健康记录的数据挖掘

利用EHR数据的挑战		
理解困难	从EHR数据中不能直接理解数据规律	序列数据预测
标签稀疏	检验的结果一般不会体现在报告上	自监督学习
缺失严重	患者不会每次进行所有检查	异构数据&数据生成
风险极高	误诊风险与隐私风险极大地限制了预测工作	隐私保护&可解释性

使用异构数据的电子健康记录预测方法

数据集	妊娠高血压预测	住院死亡率预测
样本总数	48,586	20,378
正样本数量	3,583	2,610
负样本数量	45,003	17,768
就诊次数	316,398	1,549,300
平均就诊序列长度	6.51	76.03

- 使用目标数据和辅助数据预训练模型
- 使用目标数据微调模型

使用生成数据的电子健康记录预测方法

- 使用真实数据训练对抗生成网络，并生成数据
- 使用真实数据以及生成数据预训练模型
- 使用真实数据微调模型

将可解释性融入电子健康记录预测方法中

- EHR+NLP语言模型
  - 将语言模型融入EHR数据的预测中，在预测过程中提供语言说明
- EHR+推荐系统
  - 将EHR数据融入可解释推荐系统中，对推荐的原因进行解释

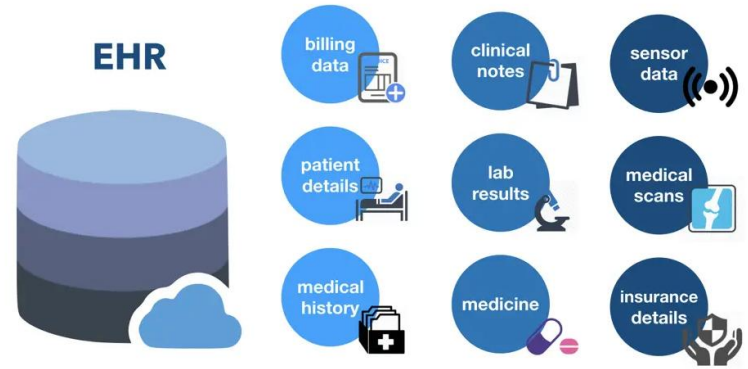
第三方图片来源

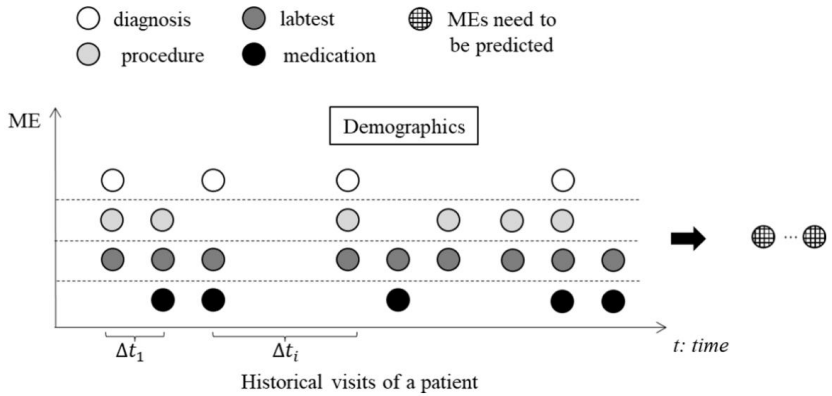
- 感谢以下作者提供的图片：
  - P2：来自twitter.com/YannR 根据CC BY-NC-SA 2.0 许可
  - P3左图：来自https://www.jianshu.com/p/79bc325379c3
  - P3右图：来自https://news.bjd.com.cn/2022/01/12/10029356.shtml
  - P4：自采集于大庆油田总医院
  - P5、8: 来自页面下方所标注论文


eClinicalWorks Blog
eCW Podcast
PRISMA
One Patient, One Record
Series
healow | Insights
WEARABLES


practice
ehr vendor
data
health system
hospital
ehr incentive
report
health record ehr
meaningful use stage
meaningful
years
health care system
information technology
technology
healthcare
medicare and medicaid
health information
clinical
physicians
system
ehr
health information exchange
medical
electronic
new
emr
providers
health information technology
medical center
electronic medical records
records
information
work
company
ehr incentive program
management
electronic health records
accountable care organizations
department of health
clinical decision support
health
privacy and security
patient
health care
patient care
care
doctors
time
services
improve


Reliability


EHR
billing data
clinical notes
sensor data
patient details
lab results
medical scans
medical history
medicine
insurance details


diagnosis
labtest
MEs need to be predicted
procedure
medication
ME
Demographics
Δt_1
Δt_i
t: time
Historical visits of a patient

谢谢

潘达岩 @ BIGSCity

dayan@buaa.edu.cn