* 1. 研究背景和科学意义

随着医疗卫生事业的发展，电子病历系统已成为医院信息系统的核心组成部分，在优化医院工作流程、提高医疗服务质量和搭建区域医疗卫生平台等方面都具有重要的作用。上世纪 90 年代，国外部分发达国家已开始建设电子病历系统 1，2010 年，我国卫生部下发了《卫生部关于开展电子病历试点工作的通知》 2，正式布局电子病历系统的建设。经过多年的发展，电子病历数据已颇具规模，可为相关研究提供新的数据平台，成为重要的国家战略资源。

近年来，人工智能飞速发展，在计算机视觉、自然语言处理、语音处理等领域都取得了巨大的突破。利用人工智能技术分析挖掘电子病历数据有助于改善医疗服务质量，具有重要的商业价值和社会意义，逐渐得到学术界和工业界的重视，许多研究工作将通用的图像处理和文本处理模型应用于医学影像和文本，实现了医学影像自动阅片 [[1](#_bookmark6)–[3](#_bookmark7)] 和文本病历实体识别 [[4](#_bookmark8)–[6](#_bookmark9)] 等功能。但目前的研究主要集中于医学影像和文本病历，针对电子医疗记录（**Electronic Medical Records, EMR**）的研究相对较少，缺少针对电子医疗记录分析的解决方案。电子医疗记录是指有固定格式的、由医生录入的、从患者入院到出院的全部记录， 包括检查、诊断、手术、用药等信息，囊括了患者一次或者多次就医的全过程信息，对电子医疗记录进行深入分析可以辅助医生进行患者管理，提升医院的诊疗服务质量，对推进智能医疗建设具有重要的意义 [[7](#_bookmark10)–[9](#_bookmark11)]。

图 [1](#_bookmark0)展示了患者就医产生电子医疗记录和基于这些数据进行分析和解释的示意图。患者在医院的每个记录可以用三元组(*recordID, value, time*) 来表示，其

1[https://www.nethealth.com/a-history-of-electronic-medical-records-infographic/](http://www.nethealth.com/a-history-of-electronic-medical-records-infographic/)

2<http://www.gov.cn/gzdt/2010-10/14/content_1722508.htm>

患者 医院





**电子医疗记录**

疾病指标



时间

N

3

Stage 4

T

2

Stage 3

T

**危险**

1

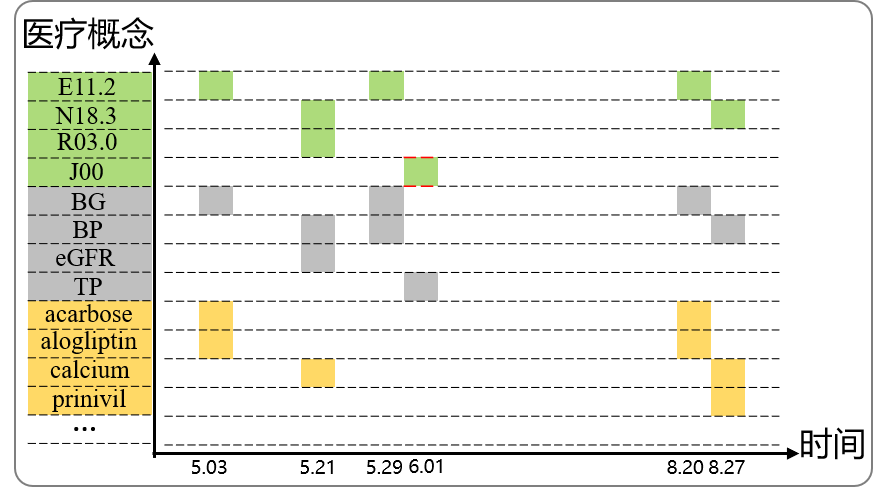
Stage 2

T

ow

Stage 1

**分析与模型解释**



① 疾病阶段预测 ③ 重入院时间

② 患者风险评估 ④ 治疗手段决策

ഴ 1 Ø$ä‰fifi⽰‰

中 *recordID* 是指医疗概念，*value* 表示该编码对应的值，*time* 表示该编码被记录的时间，以血糖为例，患者血糖被记录时会产生一条记录 (®㌌,*7.1mmol/L, 2020-02-29 15:20:32*)。*recordID* 通常采用医院规定的标准编码，在不同地区和医院之间有一定的通用性，也是针对电子医疗记录进行分析的基本特征。患者每次就医产生多条电子医疗记录，即有多个 *recordID* 被记录下来，同时，患者可能多次访问医院，因此，如图所示，电子医疗记录可用二维矩阵表示，横轴表示时间，记录患者去就医的时间，纵轴表示不同的医疗记录编码，包括诊断、检查、手术和用药等，图中所示的患者患有第二型糖尿病（E11.2），同时伴随有慢性肾脏病第三期（N18.3）和高血压（R03.0），某次就诊时还患有感冒（J00)，就诊时医生会安排检查血压（BP）、血糖（BG）等指标，确诊后会开药，如图中所示的阿卡波糖（acarbose），阿洛列汀（alogliptin）等。可见，电子医疗记录全面记录了患者就医数据，对其进行深入分析可以辅助疾病阶段预测、患者风险评估、重入院时间和治疗方案决策等临床工作，不仅能帮助患者得到及时周到的治疗， 还能减轻医务工作者的负担，具有重要的社会意义和商业价值。

结合电子医疗记录的特点和医疗领域的应用需求，可以知道，电子医疗记录分析具有以下挑战：

1. 高维性。电子医疗记录所包含的医疗概念非常多，导致每个患者的特征维度非常高，例如：由世界卫生组织整理发布的 ICD-10 3是常用的疾病诊断

3《国际疾病与相关健康问题统计分类第十版》（the 10th revision of the International Statistical Classification of Diseases and Related Health Problems），https://icd.who.int/browse10/2016/en

编码标准，而 ICD-10 共包含 15.5 万种编码，即疾病的种类（特征数）已达10 万级规模。另一方面，相比总数 15.5 万种疾病，患者实际就诊时的疾病诊断编码只占很小的一部分，这使得如图 [1](#_bookmark0)所示的电子医疗记录是极度稀疏的。此外，电子医疗记录包含的检查（LONIC 4）、药物（NDC 5）等编码同样具有高维、稀疏的特点。

1. 不规则性。电子医疗记录因为种种原因会有大量的缺失值，如尽管医生通常会建议患者去医院的时间，但患者未必遵照医嘱，反而只会在自己不舒服时才会去医院就医。另一方面，不同医疗概念的记录频率差别较大，如血糖值需要一天内测量多次，但肌酐通常一天内只测量一次。图 [1](#_bookmark0)的示例也反映了电子医疗记录在时间维度上具有不规则性，这为直接分析电子医疗记录带来了新的挑战。值得注意的是，虽然患者就医没有固定时间，但根据观测到的患者数据，多数情况下可以推断患者的缺失数据，因为患者的健康状态并非完全随机的。
2. 易扰性。电子医疗记录的特征之间相互作用关系复杂，基于电子医疗记录构建的机器学习模型容易被混淆特征（confounder）干扰，例如：早期很多数据分析表明口服避孕药和心肌梗塞具有强关联性，但是后来研究发现，这部分节育人群种吸烟的比例较高，容易发现，吸烟这个特征混淆了口服避孕药和心肌梗塞的关系。这种特征之间的易扰性为构建可解释的医疗分析模型带来了巨大挑战，电子医疗记录分析模型的可解释性是实际应用的必要条件，只有具备足够的可解释性，才能为医务人员提供模型做出判断的原因，保证模型的可靠性和使用范围。

本项目直接面向国家医疗卫生事业和人工智能的发展战略，借鉴大数据分析的通用流程 [[10](#_bookmark12)]，基于深度学习前沿技术，研究电子医疗记录预测性分析的关键技术，从队数据获取、数据预处理和数据分析三个方面，构建面向电子医疗记录端到端的分析方案，通过解决高维性、不规则性和易扰性所带来的挑战，提升电子医疗记录分析的准确性和可解释性，为基于电子医疗记录的智能医疗服务提供核心技术支撑。本项目的研究意义主要体现在以下四个方面：

1. 本项目拟实现基于表现型（**phenotype**）的自动队列识别方法。队列识别即给定筛选条件，从电子医疗记录中选出一组患者，其中部分患者有某种疾

4https://loinc.org/

5[https://www.fda.gov/drugs/drug-approvals-and-databases/national-drug-code-directory](http://www.fda.gov/drugs/drug-approvals-and-databases/national-drug-code-directory)

病（实验组），而另外一部分没有（对照组），准确的队列识别是电子医疗记录分析的基础。但因为电子医疗记录的高维性，加之并发症、异病同治和同病异治等原因，队列识别的准确性不高。而且，传统的队列识别需要医生花费大量精力比对患者的记录和统计图表才能得到，因此难以获得大规模电子医疗记录的队列标签。表现型是指可以代表患者或者疾病特点的一组人体生化或者物理特征，通常是一组研究人员指定的队列特征。本项目拟针对电子医疗记录，构建基于表现型字典的患者表示模型，同时采用弱监督学习方法在模型中引入大规模未标注样本，提高表现型字典和患者的表示学习效果，进而实现准确的队列识别。

1. 本项目拟融合电子医疗记录缺失规律和数值特征，实现电子医疗记录缺失值插补的通用方法。将缺失部分特征值的电子医疗记录补全为完整的数据， 被称为EMR 插补。经过插补处理后的电子医疗记录可使用现有循环神经网络（Recurrent Neural Network, RNN）及其变种进行建模，能有效降低预测建模的复杂度和缓解数据不规则性带来的影响，提升预测效果。如果图 [1](#_bookmark0)所示的电子医疗记录是一个稠密矩阵，数据分析师（甚至医生）可以较容易地调用RNN 对该数据集建模，实现患者病情的预测。现有针对电子医疗记录的插补研究较少，且它们忽略了电子医疗记录中固有的医学偏差（medical bias），导致插补后的数据不够准确。本项目以特征缺失规律反映医学偏差，通过对偶RNN 模型融合缺失规律和观察到的数值特征，实现更准确、泛化能力更强的电子医疗记录插补模型。
2. 本项目拟结合特征重要性和时间关联性解释模型行为和预测结果，构建基于深度学习的可解释预测模型，助力模型落地临床应用。有别于静态数据分析，电子医疗记录的可解释性需要考虑数据中存在的时间信息，不同医疗概念的时间作用域不同，如对于糖尿病患者来说，其较高的血糖值持续出现的时间较长，而对于非糖尿病患者，血糖值虽然也可能会偏高，但持续时间较短，所以如果在多个时间点出现高血糖值，该患者更有可能患有糖尿病。由于时间维度的复杂性，现有针对电子医疗记录分析的可解释模型未充分特征的重要性和时间关联性，无法得到模型全面的解释。本项目拟结合特征重要性和时间关联性构建可解释模型，不仅可以解释给定病例的预测结果，还能总结与预测目标相关的临床路径（clincial pathway），反哺循证医学研究。
3. 本项目拟在慢性肾病进展预测和感染性休克预警两个临床任务上验证上述方案的可行性。(a) 随着糖尿病患者的增加，与糖尿病有关的慢性肾病（DKD） 患者近年来也显著增加，据研究者估计，截止 2016 年，我国至少有 2430 万DKD 患者[[11](#_bookmark13)]。本项目拟与北京大学医疗健康大数据国家研究院合作，在真实慢性肾病电子医疗记录上预测患者的病情发展情况，辅助医生对部分患者及早进行临床干预。(b) 感染性休克是急诊科常见的急危重症，流行病学研究显示，急诊患者中，7.3% 在入住 4h 内发生严重感染，12% 的严重感染患者在入住 48h 内发展为感染性休克，及时预测或发现病人的状态对降低ICU 感染性休克的病发率和死亡率具有重要意义。本项目拟在麻省理工学院公开的 ICU 数据集 MIMIC-III [[12](#_bookmark14)] 上，通过预测感染性休克验证本项目的研究成果。
   1. 国内外研究现状及发展动态分析

根据与本项目的相关性，本节从电子医疗记录的表示学习、表现型分析、数据插补、以及模型的可解释性四个方面介绍和分析国内外研究现状。

1. 电子医疗记录表示学习

电子医疗记录的表示学习借鉴于自然语言处理领域的词嵌入（word embed- ding），尤其是 word2vec [[13](#_bookmark15), [14](#_bookmark16)] 模型。表示学习的基本思想是：出现在相似上下文的特征具有相似的语义信息，即相似的向量表示。好的医疗特征表示能合理地反映医疗特征之间的语义关系，有助于电子医疗记录的分析。

电子医疗记录表示学习的相关研究工作如表 [1](#_bookmark2)所示。表中仅总结了基于电子医疗记录的特征表示学习研究成果，实际上，还有许多研究工作基于医学病历或者其他文本学习医疗术语的表示学习，但和本项目的研究对象差别较大，故不在此列举。Tran 等人 [[15](#_bookmark17)] 利用非负玻尔兹曼机（Nonnegative Restricted Boltzman Machine）学习医学特征的向量表示，是最早使用深度学习的研究方法；Miotto 等人 [[16](#_bookmark18)] 和 Choi 等人 [[17](#_bookmark19)] 也较早的利用其他深度学习模型学习学习医疗特征的表示；本项目申请人蔡祥睿 **[[18](#_bookmark20)]** 首先提出不同医疗特征的时间作用域有差别，其上下文界定不能设置为固定超参数，并引入注意力机制，学习医疗特征不同的上下文时间范围；Peng 等人[[19](#_bookmark21)] 在此基础上设计更复杂的注意力模型，进一步提高特征表示学习的效果。这些工作所使用的数据只有电子医疗记录数据集；Bai 等人 [[20](#_bookmark22)] 提出对齐多医疗机构的特征并学习其表示。Choi 等人 [[21](#_bookmark23), [22](#_bookmark24)] 率先在医疗

‡ 1 EMR ‡⽰ᆜ½ª“®nffl

|  |  |  |
| --- | --- | --- |
| 序号 | 主要思想 | 文献号 |
| 1 | 仅基于 EMR 数据集训练医疗特征表示 | [[15](#_bookmark17)] [[16](#_bookmark18)] [[17](#_bookmark19)] [[18](#_bookmark20)] [[19](#_bookmark21)] [[20](#_bookmark22)] |
| 2 | 结合 EMR 数据集和医疗领域知识库 | [[21](#_bookmark23)] [[22](#_bookmark24)] [[23](#_bookmark25)] [[24](#_bookmark26)] |
| 3 | 基于异构医疗数据学习医疗特征表示 | [[25](#_bookmark27)] [[26](#_bookmark28)] [[27](#_bookmark29)] |
| 4 | 学习层次化的 EMR 数据表示 | [[28](#_bookmark30)] [[29](#_bookmark31)] |

特征表示学习模型中引入领域知识；Ma 等人 [[23](#_bookmark25)] 也提出了相似的思路，并将所学特征表示用于疾病诊断的预测；Song 等人 [[24](#_bookmark26)] 面向多个机构的电子医疗记录表示学习，通过引入了医学领域的知识库进行语义对齐。部分研究工作通过结合电子医疗记录、病历和文献等异构数据进行模型训练，期望得到更好的特征表示 [[25](#_bookmark27)–[27](#_bookmark29)]. 此外，Choi 等人还提出了针对医疗特征和患者就诊两个层次的表示学习模型[[28](#_bookmark30)]，以及利用图神经网络学习医疗特征之间的关联关系 [[29](#_bookmark31)]，这些工作在医疗特征表示学习中引入了数据的结构知识。

Ø$ä‰fifiØfl¼¾ੇø‡⽰fl‡4ශA᧛ª“,ᱥk½8¾¾Ø㾷® n䰤从S~,fl¼¾ੇø‡⽰ᱥ‡4ශ‡⽰ek½8¾¾ºªo¼4$fflS 㾷®ªflKfiä‰fl¼¾ੇø‡⽰ᆜ½®n,ቇ8fflC‡ [1](#_bookmark2)g 4 ㊱Pflfl² ä‰fl¼¾㔉fi,¼Kªቇfi儎ª⅗‡4ශ¾‡⽰ᆜ½,“‰߼⺤8¾ᛙ@k

½o$ 从‰fl²‡4ශੇøzޮ,ሼᛙ@‡⽰fi;K儎ª⅗‡4ශ¾¾¼o

1. 电子医疗记录表现型分析

表现型是一组患者或者疾病的属性，根据表现型不同可以将患者划分成更有意义的分组，与本项目研究内容队列识别直接相关。面向电子医疗记录的表现型分析的主要方法如表 [2](#_bookmark3)所示。

针对电子医疗记录进行表现型分析的人物最早由现在 UIUC 的 Jimeng Sun 教授团队引入计算机领域，他们将电子医疗记录看作张量，利用张量分解进行降维分析[[30](#_bookmark32)]，随后，该团队还先后将基于张量分解的模型扩展到联邦学习 [[31](#_bookmark33)] 和无监督学习 [[32](#_bookmark34)] 的场景中，由于电子医疗记录的高维性，张量分解的效率较低，

Heano 等人[[33](#_bookmark35)] 和 He 等人[[34](#_bookmark36)] 研究如何利用并行和分布式技术进行加速。基于张量分解的模型缺乏对记录时间的考虑，部分研究工作将电子医疗记录建模成图， 图中的点为医疗特征，特征记录如果相邻，则给图增加一条对应的边，边的权

‡ 2 Kfi EMR ¾‡4ශzflS‰

|  |  |  |
| --- | --- | --- |
| 序号 | 主要思想 | 文献号 |
| 1 | 以张量表示 EMR 数据集，进行张量分解 | [[30](#_bookmark32)] [[31](#_bookmark33)] [[32](#_bookmark34)] [[33](#_bookmark35)] [[34](#_bookmark36)] [[44](#_bookmark46)] |
| 2 | 将 EMR 建模成图，利用图算法分析表现型 | [[35](#_bookmark37)] [[36](#_bookmark38)] [[37](#_bookmark39)] |
| 3 | 基于深度学习模型，多为监督学习 | [[38](#_bookmark40)] [[39](#_bookmark41)] [[40](#_bookmark42)] [[45](#_bookmark47)] [[41](#_bookmark43)] [[42](#_bookmark44)] [[43](#_bookmark45)] |

重与记录时间间隔相关，最后利用图算法分析患者的表现型，如图分解 [[35](#_bookmark37)]、频繁子图挖掘 [[36](#_bookmark38), [37](#_bookmark39)] 等方法。深度学习模型擅长表示学习，非常适合表现型分析， 同时，挖掘表现型也能提高深度学习模型的可解释性 [[38](#_bookmark40)]。Che 等人 [[39](#_bookmark41)] 较早的将医疗编码的层次结构引入基于深度学习的表现型分析模型；Beaulieu-Jones 等人[[40](#_bookmark42)] 提出基于去噪自编码器的半监督模型，不需要大量标注样本，是本项目的

重要参考；Baytas 等人[[41](#_bookmark43)] 利用表现型对患者进行细粒度的划分；Fu 等人[[42](#_bookmark44)] 首次提出面向表现型的深度字典学习，为本项目引入表现型字典提供了理论依据； Seymour 等人[[43](#_bookmark45)] 利用深度学习研究感染性休克的表现型，是本项目研究感染性休克预测的重要参考。

4$º†␧ᓜᆜ½¾‡4ශzfl¾ශ ł;8ªᱥ$ⴇp¾ශ,䴶㾷łø

¾ḽṭ$,™flflAᇔä‰KᲥØ ),$ffiD‰fl²‡4ශzޮ,flሼᛙ@

¾‡4ශ〶⯅!,¾1㜳Ø߼⺤¾™zᛙ@¾㊱¾,¼儎k½8¾¾߼⺤⦽,䘏ᱥ$ffiD¾ࡑ᯦8S~oª8 ᝅ¾ᱥ,䜞z®nfflA᧛º†ä‰fl¼‡⽰

¾㺂k½8¾,® [[46](#_bookmark48), [47](#_bookmark49)],Ø†Ø$ä‰fifi¾儎㔪½,䘏㊱S‰9fl8½ª ղᓜ䖹儎¾ᛙ@‡⽰,™¼†k½8¾o

1. 电子医疗记录插补

因为患者就医时间没有规律，电子医疗记录具有时间不规则性，导致针对电子医疗记录的分析模型难以取得较好的效果，通过插补进行数据预处理对电子医疗记录的分析具有重要意义。关于电子医疗记录插补的相关工作总结如表 [3](#_bookmark4)所示。

针对数据缺失值的插补很早就被关注，最传统的方式是利用均值、中位数进行插补，但缺乏对数据分布的考虑，效果较差。基于传统机器学习模型进行插补近年来也一直有研究者在关注，Beaulieu-Jones 等人 [[48](#_bookmark50)] 在电子医疗记录上对比了 12 中传统机器学习模型方法，发现 MICE [[49](#_bookmark51)] 很多情况下能取得较好的

‡ 3 EMR ᨈfiª“®nffl

|  |  |  |
| --- | --- | --- |
| 序号 | 主要思想 | 文献号 |
| 1 | 传统机器学习模型 | [[50](#_bookmark52)] [[48](#_bookmark50)] [[60](#_bookmark62)] [[61](#_bookmark63)] [[49](#_bookmark51)] |
| 2 | 基于 RNN 的缺失值预测模型 | [[51](#_bookmark53)] [[53](#_bookmark55)] [[52](#_bookmark54)] |
| 3 | 基于 AutoEncoder 模型重构缺失值 | [[54](#_bookmark56)] [[55](#_bookmark57)] |
| 4 | 基于生成模型进行缺失值生成 | [[56](#_bookmark58)] [[57](#_bookmark59)] [[58](#_bookmark60)] [[59](#_bookmark61)] |

结果。值得注意的是，Zheng 等人 [[50](#_bookmark52)] 较早的提出了要考虑电子医疗记录中的医学偏差，并利用隐马尔可夫模型加入对偏差的考虑，但该工作只推断患者的状态，不能得到缺失值的插补。随着深度学习的发展，越来越多的研究工作开始利用深度学习模型处理电子医疗记录中的缺失值，Che 等人 [[51](#_bookmark53)] 较早地利用 RNN 模型进行电子医疗记录插补，为考虑时间不规则性的影响，他们设计了时间衰减因子,Cao 等人 [[52](#_bookmark54)] 和 Suo 等人 [[53](#_bookmark55)] 构建更复杂的 RNN 模型，进一步提升 [[51](#_bookmark53)] 的效果。除了 RNN 模型，少量研究工作也以 AutoEncoder 为主干模型，实现电子医疗记录插补 [[54](#_bookmark56), [55](#_bookmark57)]。最近，本项目申请人首先提出将插补问题看成数据生成问题，并构建基于生成对抗网络（**Generative Adversarial Network**，**GAN**）的数据插补模型 **[[56](#_bookmark58), [57](#_bookmark59)]**，在该领域取得了国际领先结果。此外，Yoon 等人 [[58](#_bookmark60)] 同样利用 GAN 进行数据插补，Mattei 等人 [[59](#_bookmark61)] 首次在本问题上采用深度隐变量模型

（Deep Latent Variable Model），但它们的数据假设局限于完全随机缺失（Missing Completely At Random）和非时序数据，无法应用于电子医疗记录。

电子医疗记录中的偏差是医学领域中的常见问题，已经被许多医学研究证实：Pivovarovd 等人[[62](#_bookmark64)] 利用特征出现的频率识别临床检测数据中的偏差，Phelen 等人 [[63](#_bookmark65)] 举例说明了电子医疗记录中医学偏差产生的原因，Agniel 等人 [[64](#_bookmark66)] 设计

回顾性分析实验，证明了医学偏差在电子医疗记录中普遍存在，Vassy 等人 [[65](#_bookmark67)] 利用可视化分析发现偏差。这些研究结果指出，医学偏差可以作为一种特征，有助于理解数据。

$ffiD‰ሼäᆜᐤflfiᛙ@z$¾fl¼,‰‰ሼ¾fl²Ø$ä‰fifiᨈ fi¾ශ,ᇔ4䳆fi“¾㺂 ㋈ᓜᛙ@‰z,¼儎8%ᨈfi¾߼⺤½oDࢃC$® nªä½äfl²äᆜ ᐤ㜳¼儎¾ශ᭾Q,¼ቐ$$ª“®nflº†␧ᓜᆜ½

¾ᨈfi¾ශØfl²äᆜᐤ,$ffiD¾®n正ªfiS Sੇ1¾グðo

‡ 4 EMR ¾ශª½䠀½ª“®nffl

|  |  |  |
| --- | --- | --- |
| 序号 | 主要思想 | 文献号 |
| 1 | 将模型看作黑盒，对预测结果进行解释（Post-hoc） | [[66](#_bookmark68)] [[67](#_bookmark69)] |
| 2 | 修改模型结构，在预测同时提供可解释性（Ante-hoc） | [[68](#_bookmark70)] [[69](#_bookmark71)] [[70](#_bookmark72)] [[71](#_bookmark73)] [[72](#_bookmark74)] |
| 3 | 利用可视化界面，便于理解 | [[73](#_bookmark75)] [[74](#_bookmark76)] [[75](#_bookmark77)] |

1. 模型可解释性

拥有良好的可解释性是机器学习模型实际应用的重要保证，这在医疗健康领域尤其重要。一直以来，医学和计算机交叉领域的研究人员和工程师都青睐逻辑回归和决策树等模型，因为这些模型在给出预测结果时，可以直接给出针对结果的解释和模型的行为法则，如决策树的规则和具体某个样本的预测路径。虽然目前越来越多的研究工作开始利用深度学习对电子医疗记录建模，但可解释性是深度学习的根本难题，尚无非常完备的解决的方案。

表 [4](#_bookmark5)总结了电子医疗记录分析模型可解释性的研究。现有关于电子医疗记录分析的可解释性研究主要分为两类：1）将模型看作黑盒（black-box），待模型训练完以后再对分析结果进行解释（Post-hoc）；2）改进模型，使模型本身具有可解释性（Ante-hoc）。除了这两类模型方面的研究意外，也有不少工作研究可视化交互界面，建立人机互动的平台。LIME [[76](#_bookmark78)] 最早提出将模型看作黑盒，并逐

样本解释预测结果的方法；Panigutti 等人[[66](#_bookmark68)] 首先利用这种方法解释电子医疗记录分析模型，但该工作解释的模型缺少时间属性；随后，他们由针对考虑时间的模型应用相同的思想，提出了 Doctor XAI [[67](#_bookmark69)]。这类方法的核心思想是利用给定样本周边的样本训练一个简单模型，以此对原模型预测的可靠性做出解释，实际上这类方法无法给出原模型决策的原因。我们知道，对抗样本仅对原样本做了一点微小的改变，却能改变模型预测结果 [[77](#_bookmark79)]，因此 Post-hoc 的方法难以在医疗领域广泛应用。

另一类 Ante-hoc 的方法，主要是通过注意力机制对模型进行改造，使其具有可解释性。Choi [[68](#_bookmark70)] 于 2016 年提出RETAIN 模型，将 RNN 隐含层的部分计算转移给注意力机制，虽然一定程度上提高了模型可解释性，但其预测准确率较低；Bai 等人[[70](#_bookmark72)] 和 Ma 等人[[72](#_bookmark74)] 分别提出了针对RETAIN 模型的改进。Dipole 模型[[69](#_bookmark71)] 只能提供关于预测目标比较的重要就医记录（一次就医的所有记录），无法

进一步细粒度分析。这些模型仅考虑了动态特征，忽略了静态特征的影响。Gao 等人[[71](#_bookmark73)] 虽然同时考虑了静态特征和动态特征，但两组特征的重要性是分别计算的，没有得到整体的权重。另外，所考虑的可解释性局限于具体记录与预测目标的关联性，未考虑模型的行为，即无法简单的告诉医生模型的行为准则，每个病例都需要医生仔细分辨。

在可视化医疗分析平台方面，Kwon 等人 [[73](#_bookmark75)] 为 RETAIN 做了可视化界面； Jin 等人 [[74](#_bookmark76)] 提出了交互式的智能医疗决策辅助平台 CarePre；而 Guo 等人 [[75](#_bookmark77)] 聚焦于可视化电子医疗记录，为基于不同队列的对比研究提供辅助。这些可视化分析方法和平台，是本项目系统研发的重要参考。

䢪†4$ª½䠀½S®¾ffl“‰fi¾ශ¾人ë㺂fi¾㺂ᙱ!zfl,$ffi D‰¼MªⓥⓆ¾Ø$ä‰fifizfl¾ශ, ¾ශªä¼¾“䖹ᘷfl¼e€ ᘷfl¼fi人ë㔉Q¾䍗ä,䘎㜳KⒸ!1zflfl¼¾਎!fi¾ශ人ë¾ᖧ ,fi

¾ශ¼ב»®¾½䠀o$ffiD‰®“ª㿼!⮂®o4¾ශª½䠀zfl㔉Qo

* 1. 研究内容

本项目面向医疗卫生行业智能化需求，针对电子医疗记录分析存在的研究队列识别困难、数据不规则且稀疏、模型解释匮乏等问题，以深度学习为基础手段，研究电子医疗记录预测性分析的理论和方法，力争构建端到端的电子医疗记录分析方案，突破队列识别、EMR 插补和可解释性分析模型的关键技术，并在基于电子医疗记录的实际临床任务上验证本项目的研究成果。

项目研究工作从队列识别、EMR 插补、可解释分析模型和两个预测任务四个层次展开，本项目的挑战、科学问题和研究内容关系如图 [2](#_bookmark84)所示。各部分研究内容具体介绍如下：

1. 基于表现型的自动队列识别

队列识别（cohort identification）通常时通过预先设定患者表现型（phenotype），然后筛选符合该表现型的患者。表现型是指一组可以被观察到的人体器官物理 或者生化的指标，比如某种疾病、血型、身高等。基于表现型筛选符合条件的患者队列，是电子医疗记录分析的基础，对后续分析具有重要的意义。由于电子医疗记录特征（医疗概念）维度高，不同类型的特征来源于不同的数据表或者数据库，且许多疾病（如慢性病）的记录时间跨度很大，因此，依靠医务人员人为筛

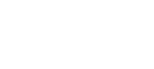


给定筛选条件，如何从EMR中识别符合条件的患者队列？

**高维性**：EMR 数据集特征纬度高，难以筛选研究队列

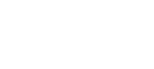


挑战



科学问题

ഴ 2 ᥇¾.‰ᆜ䰤从e®n$9“fiഴ



数据获取

数据预处理

数据分析

结合特征重要性和时间关联性的可解释预测模型

融合医学偏差的

EMR自动插补

临床任务验证

1. 糖尿病相关慢性肾病患者病情进展预测
2. 重症监护室感染性休克预警

基于表现型的自动

队列识别

研究内容

如何挖掘特征重要性和回溯分析结果， 并为医务人员提供

模型解释？

**易扰性**：特征之间相互干扰， 临床应用必须 进行模型解释

基于已经观察到的EMR，如何推断患者没来医院时的健康状态？

**不规则性**：电子医疗记录时 间维度不规律， 建模分析困难

选队列十分费时费力。

图 [3](#_bookmark85)展示了队列识别的研究内容。给定一个电子医疗数据集，包含 *N* 个患者的电子医疗记录，第 *i* 个患者表示为 ***X****i* = [***x***(1); ***x***(2); *· · ·* ; ***x***(*ti*)]，其中 ***x***(*k*) 表

*i*

*i*

*i*

*i*

示患者在时间点(*k*) 访问医院，*ti* 表示第 *i* 个患者来医院的时间点总数。需要注意的是，在相同时刻，患者会有多条记录（多个医疗特征），***x***(*k*) 的维度等于全部医疗特征念集合的元素个数。对于患者 *i* 而言，如果在对应的时间点有关于某个特征的记录，则向量对应的元素为记录值，若该特征为离散特征，则用该值为特征的取值编号或计数，若为连续特征，该值为特征的取值；若某个特征没有观测值，则 ***x***(*k*) 对应的元素没有值，记为 *NA*。预设的表现型 ***y****i* 即为本问题的标签，表现型通常具有组合性，即 ***y****i* 包含多个预定的筛选条件。对于拥有 *c* 个条件的情况，***y****i ∈ {*0*,* 1*}c*，即 ***y****i* 是一个 *c* 维向量，每一维对应一个队列筛选的条件。队列识别即给定患者的电子医疗记录 ***X****i*，判断该患者是否满足预定的表现型 ***y****i*。特征的高维性和表现型的组合性为队列识别带来了巨大的挑战，本项目首先研究针对长时间以来的层次表示模型，在不损失电子医疗记录语义的情况下，自动学习患者电子医疗记录的抽象表示。

*i*

*i*

另一方面，为训练模型实现准确的队列识别，通常需要大量的标注数据，即标注某条电子医疗记录是否满足队列预定条件，但聘请医务人员或者专家来标注成本太高，而非医学专业人员又缺乏领域知识，标注准确性较差。一般地，假设

在给定的电子医疗记录中，*M* 个样本是已经请专家标注的，记为 *L* ≜ (***X****i,* ***y****i*)*M* ，

*i*=1

而剩下的 *N* 个样本为未标注样本：*U* ≜ (***X****j,* )*N* ，通常未标注数据远多于标注

*j*=1

(𝑿𝑗, )𝑁

医疗特征

时间

(𝑿𝑖, 𝒚𝑖)𝑀

𝑖=1

𝑿

患者表示学习

无标签

数据

𝑗=1

𝒚𝑖

有标签数据

表现型

(*c* 个条件)

ഴ 3 k½8¾®n$9

数据，即 *M ≪ N* 。现有方法可分为两类：一是只是用标注样本建模，由于为充分利用未标注样本，其准确性和泛化能力较差；二是利用半监督方式训练，但在半监督模型中，标注数据和未标注数据是分开训练的，模型训练结果容易受未标注数据影响而学习到有偏的患者表示。因此，本项目另一个研究内容是在仅有少量标注数据的条件下，如何有效结合未标注数据进行训练，提升电子医疗记录表示学习效果，提高队列识别的准确率。。

1. 融合医学偏差的 **EMR** 自动插补

由于患者通常不会遵照医嘱按时就医，只有在自己不舒服的时候才会去医院，而且不同医疗特征的记录频率也差别很大，这导致电子医疗记录具有时间不规则性和稀疏性的特点，为预测性分析带来了新的挑战。本项目拟研究通用的电子医疗记录插补方法，对电子医疗记录进行预处理，使其在时间维度上更为规则，降低后续分析的模型复杂度。

图 [4](#_bookmark86)举例说明了本项目电子医疗记录插补的研究内容。给定电子医疗记录

***X***，若电子医疗记录一共有 *d* 个特征，*t* 个时间点，则 ***X*** *∈* R*d×t*。本项目拟构建机器学习模型推断 ***X*** 中的部分缺失值，即图中 *NA* 所在的位置，最终得到数值完整的电子医疗记录，如图中 ***X****ipt* 所示。要推断图中的 *NA*，可以从两个方面分析。首先，同一个特征在不同时间点的取值有相关性；其次，不同特征之间也有相互关系。可以利用时间相关性和特征之间的相关性进行建模，推断缺失值。

现有工作利用 RNN、Transformer [[78](#_bookmark80)] 和它们的变种针对时间维度进行建模， 但现有工作没有考虑电子医疗记录产生过程引入的医学偏差，导致部分插补不够准确。电子医疗记录不能直接反映患者的健康状态，因为这些记录的产生除了源自患者的健康状态，还蕴含着患者与电子病历系统的“交互”过程，比如：







 





ഴ 4 EMR ᨈfi®n$9







***ipt***

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | | 缺失标记 |  |  | 时间点 |  | |
| 1 | 0 |  0*T* | 1 |  | 插补值 |
| ***M***  0 | 1 |  1 | ***T***  3 |  |  |
| 1 | 0 |  1 | 8 |  |  |
| 原始数据 |  |  |  |  | 融合医学偏差 | 插补数据 |  |  |
|  2 | *NA* |  *NA**T* |  |  |  | 2 | 7 |  11*T* |
| ***X***  *NA* | 3 |  9  |  |  |  | ***X***  2 | 3 |  9    |
|  5 | 11 |  20  |  |  |  | 5 | 11 |  20 |

患者白细胞数目增加，但直到患者就医，且医生做出检查得到检查结果后，系统中才会记录患者的状态。由于这种数据记录过程，电子医疗记录不能直接反映患者健康状态。模型如果不能充分考虑由记录过程引入的医学偏差，容易得到不正确的结果，如患者近三年每年冬天都发生呼吸道感染，去医院检查时，发现白细胞数高于正常值，传统模型会根据多次较高的白细胞数目，推断患者在其他时间的白细胞数目也偏高，而通常呼吸道感染是短期急性病，患者康复后白细胞数目也会恢复正常。

患者就医信息被的记录过程所产生带来的医学偏差有助于电子医疗记录插补，医学研究表明 [[64](#_bookmark66)]，医学偏差不应被看作数据质量问题或者噪声，而是一种数据细分的信号，如：肾衰竭的患者更可能在晚上 10 点到早上 6 点之间化验肌酐；重病患者做检查更加频繁等。记录产生的时间可以反映特征记录的过程，是模型理解医学偏差的重要途经。如图 [4](#_bookmark86)所示，本项目拟结合患者的缺失标记矩阵和就医时间学习特征缺失规律，在 EMR 插补模型中引入医学偏差。若时间 (*k*) 时患者第 *i* 个特征被记录下来，则图中缺失标记矩阵 ***M*** 的对应取值 *m*(*k*) = 1， 若没观察到，则 *m*(*k*) = 0，缺失矩阵和记录时间完整的反映了每个医疗特征记录的时间。本项目拟研究如何利用缺失矩阵和记录时间为 **EMR** 插补模型引入医学偏差，以实现隐式的数据细分，提升 **EMR** 插补的准确性和合理性。

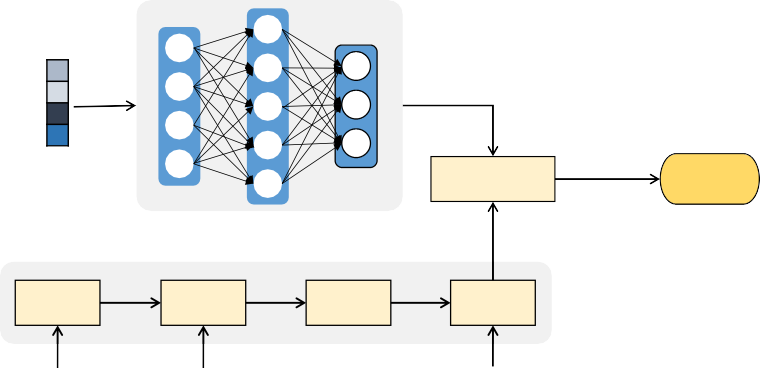
*i*

*i*

1. 结合特征重要性和时间关联性的可解释预测模型

如图 [5](#_bookmark87)所示，电子医疗记录除了包含随时间变化的医疗特征，还会包含患者的人口统计学特征，如性别、种族等，这些特征虽然是静态不变的，但对预测患者病情发展也非常重要，所以在构建预测模型时，首先应该研究如何将患者的静态特征和动态特征分别建模，以及如何将两组特征组合起来进行预测。

**MLP** 静态特征重要性



静态特征

联合表示 预测

**RNN**

**EMR**

hidden

***x***1

|  |
| --- |
|  |
|  |
|  |

hidden

***x***2

|  |
| --- |
|  |
|  |
|  |

...

hidden

***x****t*

|  |
| --- |
|  |
|  |
|  |

时间关联性

动态特征重要性

ഴ 5 ª½䠀人ë¾ශ®n$9

模型的临床应用需要提供对模型和预测结果的解释，因此，预测模型的可解释性是本项目另一个重要的研究内容。目前针对电子医疗记录预测分析的深度模型可分为三类：第一类研究没有考虑可解释性，仅在测试集上取得较高的预测准确率，这类模型不能为医生提供决策依据，不能直接应用于临床场景；第二类研究在预测模型建模时没有考虑其解释性，但利用后解释（post-interpretable） 模型来解释预测结果，这类方法具有两个缺点，一是后解释模型把任何预测模型当作黑盒，依然无法让医生了解模型真实的决策过程，只能提供预测结果的可靠程度，二是随着就诊数据的变化，需要重新训练预测模型和后解释模型，带来了额外开销；第三类研究在预测模型中实现可解释性，在完成预测的同时，也能为预测结果提供可解释的依据，这类方法尝试在预测准确性和可解释性两方面找到一种平衡，通常在模型实现可解释性时，其准确率会有所下降。

本项目拟在预测模型中考虑可解释性，现有工作缺少对静态特征和动态特征的统一比较，同时无法追踪模型的行为。静态特征通常用浅层全连接层建模， 本项目拟采用逐层相关性传播 [[79](#_bookmark81)] 分析静态特征与预测结果的关系。而对动态特征，本项目拟采用 RNN 模型进行建模，但由于 RNN 复用隐含层，其针对多维特征的可解释特别差，一方面，隐含层是由所有特征经过多步计算得到的，无法从隐含层解耦出每个特征对预测结果的贡献；另一方面，RNN 中最后的隐含层是由多步计算得到，其时间信息也无法逆向得到，阻碍了特征出现时间的分析。

所以动态特征可解释性研究的重点在于分析高维的动态特征与预测目标的关系。本项目通过对 **RNN** 隐含层解耦，挖掘动态特征重要性和时间关联性，并统一比较动态特征和静态特征的贡献，实现模型的全局解释，且可回溯样本预测结果， 为医务人员提供全面的模型解释。

1. 临床预测性任务

本项目拟将研究成果应用于具体的临床预测性任务，以验证其有效性，同时，通过和临床医生的合作，提高研究质量，加速研究成果应用。具体而言，本项目拟将研究成果应用于慢性肾脏病进展预测和急诊科感染性休克预警，并通过和医生的交流，保证项目研究内容具有实际应用价值。

* 与北京大学医疗健康大数据国家研究院合作，预测 **DKD** 慢性肾病患者的病情发展情况，提醒医生及时进行临床干预。
* 利用机器学习手段预测病人状态对降低 ICU 感染性休克的病发率和死亡率具有重要意义。本项目拟基于 **MIMIC-III [[12](#_bookmark14)]**，利用项目研究成果实现 **ICU** 患者感染性休克预警系统，验证项目研究成果的有效性。
  1. 研究目标

本项目从医疗卫生行业和人工智能结合的实际需求出发，针对队列识别泛化能力差、EMR 插补准确率低、以及预测模型可解释性匮乏等问题，研究队列识别、EMR 插补和可解释预测模型等问题，并将相关研究成果应用实际临床数据，检验本项目的研究成果在电子医疗记录预测性分析中的应用效果，推动电子医疗记录分析的研究进展。

具体研究目标包括：

在技术方面，本项目拟在以下四方面实现技术突破：(a) 提出基于表现型的队列识别方法，充分利用已标注和未标注数据，提高队列识别模型的准确性和表示学习的泛化能力；(b) 提出融合医学偏差的 EMR 自动插补方法，实现数据预处理，通过合理引入医学偏差，提高 EMR 插补的准确性；(c) 提出可解释预测模型，在进行预测的同时，可以从模型中得到与预测目标相关的特征和时间点；

1. 将本项目的研究成果应用于 DKD 患者病情进展预测和重症监护室患者感染性休克预测。

在成果形式方面，本项目力争在国内外高水平期刊、会议上发表论文 8 篇以上，全部被 SCI/EI 检索，其中有重要影响的论文 4 篇以上；申请专利 2 项。

在人才培养方面，通过本项目研究，培养深度学习、电子医疗记录分析、可解释人工智能等交叉领域的青年人才，拟培养研究生 3-4 人。

* 1. 拟解决的关键科学问题

基于上述研究目标和研究内容，本项目拟在以下几个关键理论和技术问题上有所突破：

1. 在弱监督条件下构建泛化性强的患者层次表示，实现自动队列识别

从海量、高维的电子医疗记录中识别研究队列是电子医疗记录分析的基础， 由于电子医疗记录的高维性和表现型的组合性，依靠医务人员人力筛选和标注所有符合预定表现型的队列数据不具有可操作性。患者的表现型表示可以更准确地对应预定义表现型，提高队列识别准确度，但仅基于少量队列标签的 EMR 数据集不足以让模型学习到好的数据分布，同时，传统半监督方法在标注数据和未标注数据上分别训练，模型的训练更容易被未标注数据影响。因此，如何合理利用大量未标注数据辅助建模，构建泛化性强的患者层次表示，实现电子医疗记录的队列识别是本项目拟解决一个关键科学问题。

1. 建立医学偏差和缺失值的对应关系，并将其融合到电子医疗记录插补中

电子医疗记录具有时间不规则性，直接在电子医疗记录上建模和训练难以得到较好的结果，训练得到的模型通常容易过拟合，泛化性较差。在构建预测模型之前对电子医疗记录进行插补具有重要意义，可以降低预测模型复杂度和训练难度。现有方法忽略了电子医疗记录中固有的医学偏差，将其中的缺失值当作完全随机缺失对待，其插补结果不够准确。医学偏差并不是数据质量问题或者噪声，反而对充分理解电子医疗记录非常重要。如何建立医学偏差和记录缺失的对应关系，并将其融合到电子医疗记录插补中是本项目构建通用电子医疗记录处理框架要解决的一个关键科学问题。

1. 电子医疗记录分析模型可解释性的具象化和实现方法

模型的可解释性对医学领域的预测性分析至关重要，决定了模型是否能应用于临床实践。电子医疗记录（动态特征）是一种特殊的多元时序数据，通常采用 RNN 模型建模，但在 RNN 学习过程中，多个特征在不同时间点的取值都被糅合在RNN 的隐含层中，无法总结得到特征对于模型的意义，也无法计算出某条记录和预测结果的相关性。因此，如何具象化电子医疗记录分析模型的可解

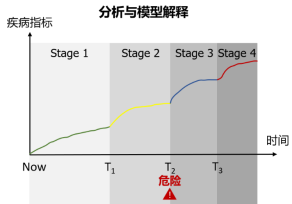
释性需求，并构建可解释预测模型，支持模型的全局解释和预测结果的回溯，也是本项目拟解决的一个关键科学问题。

* 1. 研究方案和技术路线

基于深度学习的端到端的EMR分析

数据获取 数据预处理 数据分析

专家反馈



**队列识别**

**EMR 插补**

**可解释模型**

EMR数据集

表现型字典学习

不规则时间建模

模型可解释性

弱监督学习 融合医学偏差 结果可解释性

ഴ 6 ᙱ!KfiⒸ㓵

围绕[2.1](#_bookmark83)规划的研究内容和[2.2](#_bookmark88)制定的研究目标，本项目拟定的总体技术路线如图 [6](#_bookmark89)所示。本项目针对电子医疗记录分析构建基于深度学习的端到端的解决方案，通过对电子医疗记录数据集依次进行队列识别、EMR 插补和可解释预测， 实现对特定患者队列的分析挖掘。本项目的主要研究内容分别对应大数据分析通用流程中的数据获取、数据预处理和数据分析。此外，鉴于医疗领域的特殊性，本项目的研究工作需要与医生进行充分的沟通，在模型设计和结果解释等方面融入医生的反馈。下面针对各部分研究内容，详细介绍其具体研究方案和技术路线。

1. 基于表现型的自动队列识别

要实现针对电子医疗记录的队列识别，首先需要构建患者的电子医疗记录表示学习模型，由于电子医疗记录特征维度高，本项目拟采用层次表示学习模型进行建模，同时，在模型中充分利用未标注数据，提高模型的学习效果。

图 [7](#_bookmark90)展示了基于表现型的自动队列识别研究方案，如图所示，方案首先利用RNN 对电子医疗记录建模，得到患者不同时间点就医记录的向量表示。这一步有多种模型设计选择，如采用双向RNN 模型、加入注意力机制通常能得到更好表示学习结果，这对于本文电子医疗记录建模非常重要，但序列建模的研究比

***p***  *f* ({***D****i* }, ***w***)

基于表现型的患者表示

标注数据

队列识别

***y***

表现型字典***D***

𝒉1

𝒉2

𝒉𝑡

**RNN** 数据重构

**EMR *x***

1

***x***

2

***x***

*t*

***x***1 ***x***2  ***x****t* 

hidden

...

hidden

hidden



ഴ 7 º†‡4ශ¾g€k½8¾®nS¾

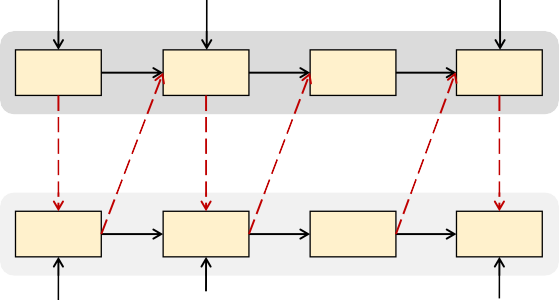
较成熟，此处不再赘述。在得到患者的基本表示 (***h***1*,* ***h***2*, . . . ,* ***h****t*) 后，以其作为上下文，在表现型字典 ***D*** 中查找相关的表现型向量（如图所示，即 *{****D****i}*），同时得到表现型字典向量重要性 ***w***，以 ***p*** = *f* (*{****D****i},* ***w***) 作为患者的表现型表示。至此，构建了 **“**特征 *→* 一次就医 *→* 表现型 *→* 患者**”** 的电子医疗记录分层表示模型。随后，对已标注数据，不仅利用基于表现型的患者表示进行分类，还输入重构任务，重构患者的电子医疗记录，而对未标注数据，则利用基于表现型的患者表示重构电子医疗记录，引导表现型字典 ***D*** 和对应的患者表示 ***p*** 学习已标注和未标注数据的整体分布。可以发现，已标注和未标注数据的任务可以一起训练。此外，由于未标注数据只需要查找表现型字典和进行数据重构，所以这部分模型参数可以用未标注数据进行预训练，可作为已标注数据训练的初始参数。

现有电子医疗记录表示学习的工作主要学习医疗特征向量和患者某次就医向量，尚没有针对表现型的向量表示学习。相比于现有工作, 本方案可对患者的向量表示进行分解，得到各个患者共享的表现型字典向量，保留患者语义信息用于重构任务。此外，本项目还将从以下两方面进一步改进模型：

* + 利用标签嵌入（**label embedding**）的方法，建立队列识别目标 ***y*** 和表现型字典的映射关系，提高了队列识别的准确性和模型的可解释性；
  + 引入医疗领域与表现型相关的知识库，如 PheKB 6，构建表现型字典和领域知识库的映射关系，并利用知识库作为模型训练的约束条件，引导表现型向量训练。

6https://phekb.org/phenotypes

**Mask *m***1 ***m***2



hidden

hidden

...

hidden

hidden

hidden

...

hidden

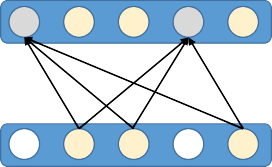
**RNN**

**RNN**

**EMR *x***1 ***x***2

***m****t*

 NA  观测值  推断值

'

***x***

*i*

***x****i*

***x****t*

（a）对偶学习 （b）特征回归

ഴ 8 㷃¼äᆜᐤ¾ EMR g€ᨈfi®nS¾

1. 融合医学偏差的 **EMR** 自动插补

融合医学偏差的 EMR 自动插补旨在将在时间维度上不规则的电子医疗记录补全为完整的电子医疗记录，可降低后续预测模型的复杂度和训练时间。电子医疗记录除了反映患者的身体健康状态，还包含的患者与医院、医生与电子病历系统的交互过程，通常这些附加因素会引入医学偏差。研究表明，医学偏差可用于对患者更精细的分类，对于理解电子医疗记录有重要辅助作用，许多医学偏差都会通过医疗特征被记录的时间推断出来，这为建模引入医学偏差提供了理论基础。本项目利用医疗特征记录的时间和电子医疗记录中缺失值出现的位置，学习特征的缺失规律，进而将所学缺失规律融入插补模型中。

图 [8](#_bookmark91) 展示了本项目EMR 插补方案的两个核心模块，首先本项目拟针对缺失标记（即图中的 (***m***1*,* ***m***2*, . . . ,* ***m****t*)）建模，如 图 [8](#_bookmark91) (a) 所示，缺失标记可以反映特征是否被记录，通过对缺失标记建模，可以得到各个医疗特征在不同时间的缺失规律，即患者医疗特征被记录的规律。同时，本项目对已经观察到的医疗记录进行建模，这也是缺失值插补最重要的推断基础。我们发现，在电子医疗记录中，缺失矩阵 (***m***1*,* ***m***2*, . . . ,* ***m****t*) 和观察到的数据 (***x***1*,* ***x***2*, . . . ,* ***x****t*) 具有对偶关系， 即缺失标记和观测数据是相互影响的，可作为互相取值推断的参考。以新冠肺炎为例，按照国家卫健委制定的诊断标准，如果患者 2 次核酸检测为阴性，则证明其已经康复，所以接下来短时间内患者的核酸检测结果将会缺失，反之，在医疗条件具备的情况下，如果患者长时间没有被安排核酸检测，则该患者很可能没有患病，核算检测结果也较大概率为阴性。因此，本项目拟通过对偶学习方法，融合缺失矩阵和数值矩阵对缺失值进行推测。

其次，电子医疗记录包含多元特征，特征之间一般具有潜在的相互关系，为插补合理的缺失值，应考虑特征之间的相互关系。如图 [8](#_bookmark91) (b) 所示，本项目拟在建模过程中引入特征回归方法，以深度学习模型对特征间关系进行建模，在统一时间点，若以观察到部分数据，可用它们推断缺失值，这种方法被称为特征回归。图 [8](#_bookmark91)

1. 展示的特征回归建模为一个线性层，可以建模为 ***x****′i* = *f* (***W*** *,* ***b***; ***x***) = ***Wxi*** + ***b***。最后，由于不同特征记录时间的差异性，电子医疗记录还具有特征层面的不

规则性，即不同特征的记录时间、记录频率差别很大，所以直接使用原始 RNN 模型进行EMR 插补容易丢失特征在时间维度方面的信息，因此，本项目拟通过修改 GRU（Gated Reccurent Unit）的门机制，在 GRU 中引入时间衰减因子，通常而言，特征记录时间离当前时间越远，对患者当前的状态影响越小，反之，特征记录时间离当前时间越近，则对患者当前的状态影响更大。通过在 GRU 中引入时间衰减因子，可以细腻度的控制特征层面的不规则性带来的影响。

1. 结合特征重要性和时间关联性的可解释预测模型

电子医疗记录中包含静态特征（如：性别、籍贯等）和动态特征（如：血糖值、慢性肾病分期等），显然，静态特征随着时间的推移不会发生变化，因此，本项目在构建预测模型时拟首先针对静态特征和动态特征分别建模，然后将两组特征的向量表示融合起来。

预测模型的可解释性对于临床应用具有重要意义，就电子医疗记录预测性分析而言，模型的可解释性可以总结为两方面：1）对于给定研究队列和模型，分析特征的重要性，称为模型的全局可解释性；2）对于某个样本的预测结果，提供导致该结果的主要因素，即一个或者一组具体时间的具体特征值。

由于电子医疗记录中的静态特征远少于动态特征，建模复杂度较低，本项目拟采取单层全连接神经网络进行建模，有助于直接计算特征的重要性。而对于单个样本预测结果的解释，本项目拟通过深度泰勒分解（**Deep Taylor Decom- position**）算法 [[80](#_bookmark82)] 分析针对静态特征的部分模型。对动态特征，本项目拟采用RNN 模型进行建模，虽然 RNN 能有效捕捉电子医疗记录中的长时间依赖的关系，但由于其隐含层多次计算，且参数复用，难以通过现有方法得到特征的重要性和具体记录的重要性。本项目拟设计混合注意力机制，解耦时序模型中特征的相互关系和时间的依赖性。

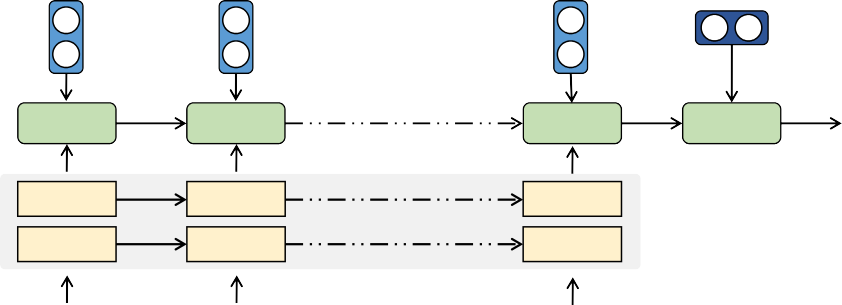
图 [9](#_bookmark92)是本项目针对动态特征的可解释性研究方案的示意图，以图中描述的 2

个特征为例，首先，为了解耦隐含层是由多个特征计算得到的，我们对输入 ***x****i*

② 特征敏感的时间注意力 ***I****T*

③ 特征注意力 ***I****F*

***y***



Attn

Attn

Attn

MLP

***h***11

***h***12

***h***21

***h***22

***h****t*1

***h****t*2

①

**RNN**

**EMR *x***1 ***x***2 ***x****t*

*f*1 *f*2

ഴ 9 ª½䠀人ë¾ශC€ᘷfl¼P®nS¾

的每个特征分别计算其隐层表示，然后以拼接的形式组成 ***x****i* 的隐含表示 ***h****i* = [***h****i*1*,* ***h****i*2]，这样分开训练虽然避免了特征组合对可解释性的影响，但这样做会降低模型预测的准确率，本方案利用混合注意力机制实现特征自动组合和转换，在提高模型可解释性时，也不影响预测性能。

本方案的混合注意力机制包括特征敏感的时间注意力和特征注意力两部分，

特征敏感的时间注意力参数为 ***I****T ∈* R*v×t*，其中 *v* 表示特征的个数，*t* 代表时间点的个数。输入 ***x****i* 的隐含层表示 ***h****i* 通过图中所示的时间注意力模块（Attn）进行特征融合，具体而言，时间注意的模块的输出为 ***I****Ti ⊙* ***h****i*，其中 *⊙* 表示沿着时

间维度利用特征隐向量乘以注意力权重，即 ***I****Ti*

*⊙* ***h****i* = ∑*v*

*ITij*

***h****ij*。通过引入

特征敏感的时间注意力权重，可学习每个特征在不同时间点（即每条记录）的重

*j*=1

要性，而特征注意力参数 ***I****F ∈* R*v* 则用于学习特征的对于模型和预测结果整体重要性，在最后的全连接层之前与静态特征表示融合并引入模型，以便实现静态特征和动态特征的统一比较。在此框架下，动态特征的可解释性可建模为以下形式：

*v*

∑

*p*(***y****|****X***) = *p*(*y|z* = *i,* ***X***)*P* (*z* = *i|****X***)

*i*=1

*v*

∑

= *p*(*y|z* = *i, IT*1*i*

*i*=1

* ***h***1*i, IT*2*i*
* ***h***2*i, . . . , ITti*
* ***h****ti*)

(1)

* + *P* (*z* = *i|IF*1 *·* ***h****t*1*, IF*2 *·* ***h****t*2 *. . . , IFv ·* ***h****tv*)

其中 ***X*** = (***x***1*,* ***x***2*, . . . ,* ***x****t*) 为模型输入，*z* 是隐变量，表示数据特征的编号，*z ∈*

1*,* 2*, . . . , v*，用于解释不同特征的重要性和时间关联性。*IT*1*i ·* ***h***1*i, IT*2*i ·* ***h***2*i, . . . , ITti ·* ***h****ti* 即为特征 *i* 时间维度的注意力，而 *P* (*z* = *i|IF*1 *·* ***h****t*1*, IF*2 *·* ***h****t*2*, . . . , IFv ·* ***h****tv*) 则为特征 *i* 的全局重要性。

1. 临床预测性任务

从 DKD 慢性肾病和急诊科感染性休克的临床需求出发，利用本项目的研究成果，研发预测分析系统：

首先，熟悉 DKD 慢性肾病和急诊科感染性休克的医学背景，与医生合作， 标注少量符合表现型的队列数据，利用第 (1) 节研发的模型从大规模电子医疗记录中筛选出研究队列；其次，对获得的研究队列进行预处理，分析记录时间分布的特点，利用第 (2) 节的研究成果对研究队列进行缺失值补充，并评估数据插补的效果；以第 (3) 节的核心方法构建可解释预测模型，评估其预测准确率，并研发特征重要性和时间关联性的可视化系统，展示模型和预测结果的可解释性分析效果。研究过程中，医生的反馈可作为本项目队列识别、数据插补和可解释模型研究的重要指导，确保本项目的研究方向正确，保证研究成果可应用于临床实践。

* 1. 可行性分析

1. 理论可行性

本项目研究目标明确，研究内容清晰，研究方案和技术路线中所应用的方法和技术手段在业界都有着成熟清晰的理论基础。申请人和项目组对这些关键技术和理论有着深入的了解和掌握，近年来在人工智能、医疗数据分析等领域的高水平会议上发表了多篇论文。项目组前期已经对本项目中提到的研究内容分别进行了详细的调研和分析，并在医疗特征表示学习、时间序列插补等领域取得了初步成果。因此，从理论上说，本项目是可行的。

1. 技术可行性

申请人前期调研了大量基于深度学习的电子医疗记录分析的研究工作（如[1.2](#_bookmark1)节所述），深度学习模型不仅能挖掘高维特征之间复杂的关系，同时能有效处理数据中长时间依赖关系，非常适合电子医疗记录分析。利用深度学习技术解决电子医疗记录预测性分析需要解决三个核心问题，即队列识别、EMR 不规则性处理和模型的可解释性，申请人之前的研究工作一直聚焦于电子医疗记录的分析，在医疗特征表示学习、多元时序数据插补等方面已取得一定的研究成果，

并在攻读博士期间与医生合作，参与过医疗分析系统的开发，相关研发经验可作为本项目的研究基础。申请人所在项目组多年来一直活跃在在大数据处理和分析、机器学习等领域，在相关领域有着丰富的研究经验和技术积累。同时，在研究过程中，申请人与医疗机构建立了合作关系，积累了大量可用于实验的真实数据集。因此，从技术上说，本项目是可行的。

1. 团队合理性

项目组在大数据分析和人工智能领域具有一定的基础，积累了丰富的研究经验，在重要国际会议上发表了多篇高水平论文，项目在信息系统和数据分析系统开发方面也有丰富的积累，可为本项目可视化分析系统研发提供保障。项目组梯队完善，队伍具有凝聚力和创造力，项目组成员每周定期讨论，有着良好的科研氛围，同时对本项目的研究内容具有浓厚的研究兴趣。申请人与联合培养时的导师新加坡国立大学教授Beng Chin Ooi（黄铭钧）一直保持密切联系，Ooi 教授长期研究大数据管理与分析，是数据库和数据挖掘方面非常活跃的科学家， 可为本项目提供技术指导。

申请人与医疗机构和专家一直保持良好的合作关系，与北京大学医疗健康大数据国家研究院张路霞教授合作研究慢性肾病患者病情进展，和天津市肿瘤医院合作研究肺癌患者治疗效果评估，同时，申请人参与建设“南开大学-基准联合医学研究中心”，和广州基准医疗有限公司合作研究基于多模态医疗数据的癌症早诊技术。这些合作单位和专家可为本项目提供专业意见的反馈，保证研究内容符合医学常识和临床需求。综上所述，项目团队组成合理，能保障本项目的顺利完成。

与现有研究工作相比，本项目的特色与创新之处体现在以下几方面：

* 1. 针对电子医疗记录的队列识别问题，本项目提出基于表现型的电子医疗记录表示学习，并提出新的弱监督学习方法，在仅有少量标注数据的情况下实现准确的队列识别。目前针对电子医疗记录的表示学习局限于对医疗特征、一次就诊和某位患者的表示学习，缺乏针对表现型的表示学习，而患者的状态通常由多个表现型组合而成；另一方面，将患者的表示分解为多个表现型向量，有利于利用未标注数据，提高表现型表示学习的质量，因此， 在大量电子医疗记录难以标注的情况下，本项目利用表现型字典表示学习进行队列识别具有重要的实际意义，是本项目一大特色和创新之处。
  2. 针对电子医疗记录不规则性的问题，本项目通过融合医疗特征记录过程引入的医学偏差，提出电子医疗记录插补方法，可作为电子医疗记录分析的通用数据预处理方法。为提高电子医疗记录的分析效率和降低模型设计的复杂度，本项目提出电子医疗记录实现缺失值插补模型，考虑电子医疗记录中存在医学偏差，且这种偏差与医疗特征记录的时间密切相关，本项目利用缺失标记矩阵自动学习各医疗特征的缺失规律，解决医学偏差为插补模型带来的影响，可有效提高电子医疗记录的插补质量，进而提高预测模型的准确性，是本项目相比于现有工作的重要创新之处。
  3. 现有电子医疗记录的分析模型可解释性不足，本项目提出针对特征重要性和时间关联性两方面进行解释，通过引入混合注意力机制，解耦类 **RNN** 模型的隐含层，实现针对电子医疗记录的可解释分析模型。可解释性对医疗领域的应用落地十分重要，也是在医疗机构应用机器学习模型的重要需求之一，现有电子医疗记录可解释性的研究较少，少量工作曾探索如何挖掘与预测结果相关性较大的记录，但缺少对模型行为的解释，即不能得到全局的特征重要性。本项目创造性地在分析记录的时间关联性的同时，总结特征的重要性，能更全面的展现模型的行为和决策理由，为在临床实践中应用奠定了重要的基础，也是本项目的重要特色与创新之处。
  4. 年度研究计划

本项目实施 3 年，具体年限为 2021 年 1 月 1 日至 2023 年 12 月 31 日。具体计划如下：

第一阶段（**2021** 年 **1** 月**–2021** 年 **12** 月），重点研究基于表现型的自动队列识别，同时请教专业医生关于 DKD 慢性肾病和感染性休克的表现型，具体包括：

1. 研究整理 DKD 慢性肾病和感染性休克的常见的表现型并标注；
2. 研究弱监督条件下的自动队列识别方法；
3. 实现针对 DKD 慢性肾病和感染性休克的队列识别；
4. 在国际期刊或会议上发表研究论文 2-3 篇，参加国际会议 1 次，并在会议上做论文成果报告。

第二阶段（**2022** 年 **1** 月**–2022** 年 **12** 月），重点研究融合医学偏差的电子医

疗记录查插补方法，并实现针对 DKD 慢性肾病和感染性休克的队列识别和数据补全，具体包括：

1. 研究融合医学偏差的电子医疗记录查插补方法；
2. 实现针对 DKD 慢性肾病和感染性休克的数据插补预处理；
3. 在国际期刊或会议上发表研究论文 2-3 篇，参加国际会议 1-2 次，并在会议上做论文成果报告；
4. 申请发明专利 1 项。

第三阶段（**2023** 年 **1** 月**–2023** 年 **12** 月），重点针对电子医疗记录分析的可解释预测模型，构建端到端的电子医疗记录分析解决方案，具体包括：

1. 构建电子医疗记录的可解释预测模型，支持特征重要性和时间关联度挖掘；
2. 针对 DKD 慢性肾病和感染性休克的构建可解释预测模型，研发可视化模型解释界面，并与医学专家讨论模型的可解释结果；
3. 在国际期刊或会议上发表研究论文 2-3 篇，参加国际会议 1 次，并在会议上做论文成果报告；
4. 申请发明专利 1 项；
5. 项目总结，完成结项报告，准备验收。
   1. 预期研究成果

本项目的预期研究成果包括以下几个方面：

* + 1. 在国内外高水平期刊、会议上发表论文 8 篇以上，全部被 SCI/EI 检索， 其中有重要影响的论文 4 篇以上；
    2. 申请专利 2 项；
    3. 培养研究生 3-4 人；
    4. 完成一个基于深度学习的端到端的电子医疗记录分析框架；
    5. 开源相关研究工作，供用户下载，并提供说明和使用文档。