分类号： R319 单位代码： 10335

密 级： 无 学 号： 21815048



硕士学位论文



**中文论文题目：基于贝叶斯隐马尔可夫模型的心衰演变过程分析方法研究**

**英文论文题目：Study on Progression Analysis of Heart Failure Based on Bayesian Hidden Markov Model**

申请人姓名： 王慧兰

指导教师： 吕旭东 教授

合作导师：

专业名称： 生物医学工程

研究方向： 医学信息学

所在学院： 生物医学工程与仪器科学学院

**论文提交日期**

**基于贝叶斯隐马尔可夫模型的心衰演变**

**过程分析方法研究**



**论文作者签名:**

**指导教师签名:**

论文评阅人1：

评阅人2：

评阅人3：

答辩委员会主席：

委员1：

委员2：

答辩日期： 年 月 日

浙江大学研究生学位论文独创性声明

本人声明所呈交的学位论文是本人在导师指导下进行的研究工作及取得的研究成果。除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得 **浙江大学** 或其他教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示谢意。

学位论文作者签名： 签字日期： 年 月 日

学位论文版权使用授权书

本学位论文作者完全了解 **浙江大学** 有权保留并向国家有关部门或机构送交本论文的复印件和磁盘，允许论文被查阅和借阅。本人授权 **浙江大学** 可以将学位论文的全部或部分内容编入有关数据库进行检索和传播，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文。

（保密的学位论文在解密后适用本授权书）

学位论文作者签名： 导师签名：

签字日期： 年 月 日 签字日期： 年 月 日

# 致 谢

时间过得张牙舞爪，光阴逃得死去活来。两年半的研究生生涯转眼就要结束了，这两年半是我成长最多的时光。熬过迷茫纠结的夜晚，也享受过温暖感动的黄昏。不管是在科研工作上还是为人处世上，都有了全新的理解和感受。虽然还有很多需要成长的地方，但希望自己能够乐观坚定地走向下一个旅程！值此之际，向所有帮助过我、鼓励过我的人致以最诚挚的谢意。

感谢段会龙老师，为我们提供了方便舒适的工作环境，让我们可以自由自在地在科研的世界中探索。段老师在学术上严谨深邃，而上课时幽默轻松，深受大家欢迎。

感谢吕旭东、黄正行两位老师对我的指导。吕老师为我打开了研究生的大门，

还记得在每周二上午的研讨会上，吕老师对每位同学的汇报进行了深刻、精彩的点评与指导，令我获益匪浅；黄正行老师对待科研事无巨细，什么工作都能做得非常完美，同时也非常关心同学们的生活，非常感谢黄老师在我迷茫时给予我的鼓励和谆谆教诲。

感谢王雪立师兄、胡丹青师兄和孙周健师兄在技术上对我的指导和帮助，开阔了我的思路，让我能够在科研工作上尽快适应。

感谢我的室友们，虽然都比我小，但是感觉自己被你们照顾着。谢谢你们包容我写论文这半年晚归打扰你们休息，谢谢你们对我的鼓励。

感谢江红娟，这两年半我经历了很多黑暗时刻，一度对自己的方方面面产生了怀疑，是你每次都安慰我、鼓励我，苦口婆心给我讲道理。我的成长很大一方面得益于你，很高兴、很荣幸有你这么个闺蜜。衷心地希望你在国外过得快乐，申上好的学校攻博，期待早日见面！

感谢我亲爱的男友黄浩策，斯人若彩虹，遇上方知有。谢谢你一直陪在我身边，陪我品尝生活的酸甜苦辣，陪我慢慢成长。谢谢你温柔耐心，包容我的任性。未来还有好长的路，我们要一起走下去！

最后感谢我自己，感谢自己在痛苦难熬时没有放弃，感谢自己一直咬牙坚持！

王慧兰

2020年12月于求是园

# 摘 要

心力衰竭等慢性病通常随着时间的推移进展缓慢，对此类慢性病的演变过程进行分析可以帮助医护人员更好地了解和分析患者的疾病进展，实现相关症状的早期发现、早期预防、个性化诊疗和精准预后，是为慢性病提供更好的全程管理的基础和关键。传统的疾病演变过程分析是通过队列研究来进行的，但这些研究往往仅针对目标疾病的几个精心挑选的风险因子，基于严格的入组和出组条件，难以纳入新发现的风险因子，且难以处理数据缺失的问题。

近年来，随着电子健康记录(Electronic Health Record, EHR)等医疗信息系统的快速发展，大量研究开始利用数据挖掘技术分析电子病历数据，进行疾病演变过程建模（Disease Progression Modeling, DPM），挖掘疾病的演变模式。相比传统的队列研究，DPM没有严格的入组和出组条件，且EHR中记录了更丰富的患者诊疗信息，便于纳入潜在的风险因子。然而，目前此类研究仍然存在以下不足：1）现有的模型往往只考虑了一种单一固定的疾病演变模式，而在真实世界中不同的患者可能有不同的演变模式和速度；2）现有模型没有充分利用EHR数据，而只是用到了其中小部分类型单一的患者特征。

针对以上研究的不足，本论文提出了一种基于贝叶斯隐马尔可夫模型（Bayesian Hidden Markov Model, BHMM）的概率生成模型，以心力衰竭（以下简称心衰）作为典型的慢性病代表，对心衰演变过程进行建模。本论文提出的模型不局限于单一的心衰演变模式，允许不同的心衰患者有不同的演变轨迹。并且训练模型的EHR数据集包含了患者的实验室检查结果、用药清单及手术记录等各类诊疗信息，充分利用了EHR数据。最后为了方便临床医护人员理解和使用模型，设计并实现了心衰数据分析系统。

将本论文提出的模型应用到从国内某三甲医院心内科收集到的2102份心衰患者的EHR数据进行实验，实验结果表明：1）相比传统的DPM，本论文提出的模型能够更好地理解和代表患者的心衰演变轨迹。2）本论文提出的模型能够以较高的准确率来预测心衰患者未来的临床表现。

**关键词**：疾病演变过程，贝叶斯隐马尔可夫模型，概率生成模型，心力衰竭，电子健康记录

# Abstract

Chronic diseases such as heart failure usually progress slowly over time. Analyzing the progression of such chronic diseases can provide healthcare practitioners better understanding of disease progression, and achieve the early detection, early prevention, personalized diagnosis and treatment and accurate prognosis of related symptoms. It is the foundation and key to provide better overall management of chronic diseases. Traditional disease progression analysis is carried out through cohort studies, but these studies only focus on a few hand-picked risk factors of the target disease, and they are usually based on strict inclusion and exclusion criteria. Moreover, they cannot incorporate newly-discovered risk factors, and lack the ability to deal with missing data.

Recently, with the rapid development of electronic health records (EHR) and other medical information systems, lots of studies have begun to exploit EHR to perform disease progression modeling (DPM) and draw the disease progression pattern through data mining. Compared with traditional cohort studies, EHR-based methods do not have strict inclusion and exclusion criteria, and they can incorporate potential risk factors easily as there is more patient infromation available in EHR. However, such researches still contain the following deficiencies: 1) Existing models only consider a single fixed disease progression pattern, but different patients in the real world may have different progression patterns and speeds; 2) Current models do not make full use of EHR data, but only incorporate a small part of the patient features of a single type.

To address the above limitations, this thesis proposes a probabilistic generative model based on Bayesian Hidden Markov Model (BHMM) to model the heart failure (HF) disease progression, taking HF as a typical chronic disease. The proposed model is not limited to a single HF disease progression pattern, and allows different HF patients to have different progression trajectories. To make full use of EHR, the dataset used to train the proposed model contains various patient information such as laboratory examination results, medication lists, operation records, etc. Finally, in order to facilitate physicians to understand and use the model, a data analysis system for HF patients has been designed and implemented.

The model proposed in this thesis was applied to an EHR dataset containing 2102 HF patients collected from the cardiology department of a Chinese hospital. The experimental results show that: 1) Compared with traditional DPM, the model proposed in this thesis can achieve a better understanding and representation of HF disease progression trajectory. 2) The proposed model can predict the future clinical performance of HF patients more accurately.

**Key Words:** Disease Progression, Bayesian Hidden Markov Model, Probabilistic Generative Model, Heart Failure, Electronic Health Record

# 目录

[致 谢 I](#_Toc59008946)

[摘 要 II](#_Toc59008947)

[Abstract III](#_Toc59008948)

[目录 V](#_Toc59008949)

[1 绪论 1](#_Toc59008950)

[1.1 引言 1](#_Toc59008951)

[1.2 研究现状 2](#_Toc59008952)

[1.2.1 队列研究 2](#_Toc59008953)

[1.2.2 基于EHR的研究 3](#_Toc59008954)

[1.3 研究目标与内容 5](#_Toc59008955)

[1.3.1 研究目标及任务 5](#_Toc59008956)

[1.3.2 论文内容安排 6](#_Toc59008957)

[2 数据分析 7](#_Toc59008958)

[2.1 基线数据统计 7](#_Toc59008959)

[2.1.1 基本信息统计 7](#_Toc59008960)

[2.1.2 实验室检查结果统计 8](#_Toc59008961)

[2.1.3 用药信息统计 10](#_Toc59008962)

[2.1.4 手术信息统计 10](#_Toc59008963)

[2.1.5 合并症统计 11](#_Toc59008964)

[2.1.6 心功能分级统计 12](#_Toc59008965)

[2.1.7 超声心动图统计 12](#_Toc59008966)

[2.2 案例分析 13](#_Toc59008967)

[2.3 本章小结 16](#_Toc59008968)

[3 基于贝叶斯隐马尔可夫模型的心衰演变过程建模 17](#_Toc59008969)

[3.1 基础知识 17](#_Toc59008970)

[3.1.1 贝叶斯网络 17](#_Toc59008971)

[3.1.2 隐马尔可夫模型 18](#_Toc59008972)

[3.2 模型构建 20](#_Toc59008973)

[3.2.1 心衰演变过程建模问题的定义 20](#_Toc59008974)

[3.2.2 心衰演变过程模型的提出 21](#_Toc59008975)

[3.3 实验和分析 25](#_Toc59008976)

[3.3.1 实验设置 25](#_Toc59008977)

[3.3.2 结果与分析 25](#_Toc59008978)

[3.4 模型评估 35](#_Toc59008979)

[3.4.1 评估方案设计 35](#_Toc59008980)

[3.4.2 结果与分析 35](#_Toc59008981)

[3.5 本章小结 36](#_Toc59008982)

[4 系统设计与实现 37](#_Toc59008983)

[4.1 整体框架 37](#_Toc59008984)

[4.2 功能模块 38](#_Toc59008985)

[4.2.1 群体统计 38](#_Toc59008986)

[4.2.2 个体详情 39](#_Toc59008987)

[4.3 系统实现 39](#_Toc59008988)

[4.4 本章小结 43](#_Toc59008989)

[5 总结与展望 46](#_Toc59008990)

[5.1 总结 46](#_Toc59008991)

[5.2 展望 47](#_Toc59008992)

[参考文献 48](#_Toc59008993)

[作者简介 53](#_Toc59008994)

# 绪论

## 引言

慢性病，也称为非传染性疾病，是指病情持续时间长、发展缓慢的疾病。慢性病的四个主要类型为：心血管疾病（如心脏病发作和中风）、癌症、慢性呼吸道疾病（如慢性阻塞性肺病和哮喘）以及糖尿病[1]。根据世界卫生组织(WHO, World Health Organization)最新发布的《2020年世界卫生统计》[2]显示，慢性病患病率近年来不断增加，引起的死亡总人数也不断增加，这种趋势遍及全球所有人群。在全球，有18%的人因心血管疾病、癌症、糖尿病、慢性呼吸系统疾病中的任一种而过早死亡。2017年，全球共有5600万人死亡，其中70%以上的人死于非传染性慢性病[3]。

心血管疾病(CVD, Cardiovascular Disease)是四大慢性病中危害最大，致死率最高的[3],[4]。心血管疾病是心脏和血管疾患引起的，包括冠心病（心脏病发作）、脑血管疾病（中风）、血压升高（高血压）、周围动脉血管疾病、风湿性心脏病、先天性心脏病和心力衰竭[5]。根据WHO报道，心血管疾病是全球的头号死因：每年死于心血管疾病的人数多于任何其它死因。2016年，估计有1790万人死于心血管疾病，占全球死亡总数的31％。其中85%死于心脏病和中风。四分之三以上的心血管疾病死亡发生在低收入和中等收入国家。非传染性疾病导致的1700万七十岁以下死亡中，有82%发生在低收入和中等收入国家，37%由心血管疾病造成[6]。

心力衰竭（HF, Heart Failure）,简称心衰，是大部分心血管疾病发展的终末阶段[7]。心衰是指由于心脏的收缩和舒张功能出现障碍，无法将静脉回心血量充分从心脏泵出，导致静脉系统血液滞留而动脉系统血液供给不足，从而引起心脏循环系统障碍，心脏排血量减少，不能满足人体各组织器官的需要[8]。心衰并不是一种独立的心血管疾病，而是各种心脏疾病发展到严重阶段的一种临床综合征。按发生的基本原理，心衰可分为收缩性心衰和舒张性心衰，收缩性心衰指心室收缩功能不全，左室射血分数（LVEF, Left Ventricular Ejection Fraction）下降，并有心衰症状；舒张性心衰指左室射血分数正常或接近正常，但有心衰症状[9],[10]。

心衰演变过程分析即对心衰的发展过程进行分析，从而深入了解心衰患者的整个治疗轨迹，帮助医护人员在早期阶段跟踪心衰的发展，有助于早期发现和预后,从而促进有效的医疗服务和更好的疾病管理[11]。同时，随着各大医院信息化建设进程的不断推进，医院的信息系统存储了大量的电子健康记录(EHR, Electronic Health Record)数据，包括大量的患者检查、用药及手术等临床诊疗数据，为心衰演变过程分析提供了丰富的数据来源[12]-[13][14][15][16]。因此，可以利用数据挖掘技术分析心衰患者的EHR数据，挖掘心衰的演变模式，为临床医护人员评估心衰的进展、开展针对性的医疗干预提供参考[17]。

## 研究现状

心血管病等慢性疾病的演变过程分析方法主要有队列研究和基于EHR数据两种。下面将分别描述：

### 队列研究

在临床医学研究中，慢性病演变过程分析研究一般建立在临床队列实验的基础上。队列实验指的是将研究人群按照是否暴露于某危险因素或暴露程度不同分为不同的组别，然后随访记录不同组别结局事件发生的情况，比较各组结局事件发生率的差异，从而判定暴露因素与结局事件之间的关联[18]。对于慢性病演变过程分析的队列实验，首先会制定入组和出组标准来纳入研究对象，然后通过长期随访患者群体，得到若干患者特征的长期纵向数据，利用多元逻辑回归等方法，筛选出几个独立的临床观察特征（风险因素），以其分值之和将患者划分为不同的危险等级，进而描绘出患者的疾病演变轨迹，并做出预后评价，指导治疗[20]-[21][22][23][24]。针对其中的入组与出组标准、患者特征及结局事件的论述如下[19]：

* 入组与出组标准：入组标准即对患者的症状、诊断、合并症及超声心动图等检查检验结果做出具体的要求，满足要求的患者纳入队列；出组标准主要是对会干扰实验结果准确性的因素进一步的排除，比如患者的既往病史及入院后所接受的治疗，符合出组标准的患者会被排除。通过明确的入组和出组标准，队列研究可以严格控制研究对象的特征和质量，保证实验结果的有效性。
* 患者特征：一组临床指标，用来量化描述入组对象。通常在初期的实验设计阶段，根据研究目的，确定需要长期随访来收集的患者数据项，如人口特征、体征、症状、用药以及检查检验相关结果。
* 结局事件：即对入组对象随访结束的标志，如死亡、再发脑梗/心梗等。

已经有大量针对各种慢性病演变过程分析的队列研究。Richard等人[20]根据美国、欧洲和南美共四个临床中心的随机对照临床实验，比较了非清髓性造血干细胞移植与疾病改良疗法对疾病进展的影响。Chen等人[21]提出了一个分类模型用来模拟心律失常性心肌病（AC），以发现一组不同的潜在发病机制，采用的队列包含60名具有不同遗传背景的AC患者。Annul等人[22]开展了一项纵向观察研究，以确定前房角宽度是否在窄角法向性青光眼的进展中起到相关作用。Eduardo等人[23]研究了帕金森病中自主神经功能障碍发作时间与疾病进展和生存的关系。Ito等人[24]提出了一个基于文献元分析的模型来描述轻中度阿尔茨海默病患者的纵向变化。

然而这些研究都具有以下所述的不足：

* 严格的入组和出组标准使得研究结果不具有普适性：几乎所有沿这方面的研究都倾向于针对某一种特定的疾病，研究目标也非常具体，如不同干预方法的治疗效果。在入组和出组条件的严格限制下，很多年龄较大或较小、同时患有多种合并症的患者不能进入队列，而在真实的临床环境中，这些情况是非常常见的。因此队列研究收集到的数据与真实世界存在一定的差异[25]。
* 患者特征项纳入不全：这些队列研究采用的是精选挑选的少量患者特征，无法纳入新发现或潜在的疾病特征。随着医学技术的发展，越来越多的潜在疾病特征不断被发现和重视，如估算肾小球滤过率（eGFR）等指标可能对心衰的发生和发展过程有一定的判别价值，但并未包含在NYHA心功能分级[26]等常见的心衰风险等级模型中。因此，有必要将更多的患者特征纳入到疾病演变过程分析中，分析它们影响疾病发展过程的权重，提高疾病演化过程建模的精确性[27],[28]。
* 无法处理数据缺失的问题：此类队列研究依赖完备的长期纵向数据集来建模，无法处理含有缺失值的患者。然而，在真实的临床环境中，患者记录中经常含有缺失值，因此限制了此类队列研究模型的使用[29]。

### 基于EHR的研究

随着信息技术在医疗领域内的广泛应用，大量的患者临床诊疗数据被记录并保存在EHR中。EHR在患者治疗过程中记录了各种类型的患者信息（例如症状，生命体征，实验室检查结果等）和治疗干预措施（例如药物，手术，检查等），因此隐藏了尚未被发现的大量关于疾病演变过程的信息[30]。从这个意义上讲，EHR具有独特的潜力，可用于建立疾病演变过程的数据挖掘或统计模型，然后通过这些模型帮助医务人员了解疾病进展并提供有效的健康服务，改善单个患者的治疗实践。相对于队列研究，由EHR驱动的模型没有严格的纳入和排除标准，使用的数据更符合临床的真实情况，也能够方便地纳入新的患者特征和排除具有缺失值的记录[31]-[32][33]。目前已经有一些由EHR驱动的疾病演变过程建模方法，例如：

Cohen等人[34]对在重症监护病房每分钟收集到的45个生理、临床和治疗变量进行分层聚类，以确定患者状态，识别重症监护病房中的基本疾病进展模式。结果表明，层次聚类技术可用于可视化复杂的多变量数据，并可为重症患者的护理提供新的见解。

Zhou等人[35],[36]提出了一个多任务学习框架，通过临床评分（ADAS-Cog和MMSE）来预测疾病的进展。对来自阿尔茨海默症神经影像学倡议（ADNI）数据集进行广泛的实验研究，表明提出模型的有效性。

Jackson等人[39]建立了一个多阶段隐马尔可夫模型，用于在连续时间内错误分类的马尔可夫过程。可同时估算疾病各个阶段之间的转移概率和阶段被错误分类的概率，并将其应用于动脉瘤筛查研究，根据主动脉直径的连续范围按严重程度对疾病进行分级。

Sukkar等人[40]提出了基于隐马尔可夫模型（HMM, Hidden Markov Model）的疾病进展模型。针对阿尔茨海默症的生物标记物，使用ADNI数据集以无监督的方式训练了HMM，结果表明，受过训练的HMM能够比当前定义的临床阶段更精细地模拟疾病进展。

然而，存在一些不足限制了这些模型的通用性。首先，尽管疾病的演变是一个连续的过程，但是观察数据只在一系列离散的时间点收集到，并且这些时间点之间的间隔也是不相等的；其次，单个患者的记录往往是不完整的，仅仅只涵盖了整个疾病演变过程的一小部分，必须将大量患者的记录“拼接”在一起来推出完整的演变过程[37]。针对这些挑战，已经有部分研究提出了新的解决方法：

Wang等人[37]提出了一种连续时间隐马尔可夫模型（CTHMM），通过合并症（如肾病、糖尿病）的发作来推断慢性阻塞性肺疾病（COPD）的进展，可以从不完整、不规则、离散的病人记录中学习完整的、连续的疾病演变轨迹。特别是，他们的工作将患者记录建模为时间序列，该时间序列可以在进展路径的任何点开始和结束。

Sun等人[38]使用了CTHMM和主题模型分析技术，对亨廷顿（HD，Huntington’s Disease）病人的发展阶段进行了划定，并且对每个阶段的主要特征进行了分析。实验结果表明他们的模型相比临床采用的分期标准涵盖了更精细的阶段划分，并且能够定量描述临床诊断前后的复杂变化。

但是，这类研究依然存在两点不足：1）仅侧重于发现目标疾病的单一演变模式，根据严重程度划分几个疾病状态，然后通过对马尔可夫过程加以约束使其仅能向前转换，即疾病状态只能从早期病情轻微的阶段转移到后期病情恶化的阶段。然而，仅根据严重程度来划分疾病

状态是不严谨的，更重要的是，大多数患者不太可能统一遵循单一固定的疾病演变模式，而是有多种情况。2）这些研究没有充分利用电子病历数据，而只是利用了其中的一小部分，如诊断信息或者根据现有评分模型选取的少量风险因子，鲜有研究将各种类型的数据综合起来构建更为全面、准确的模型。

## 研究目标与内容

### 研究目标及任务

本论文针对以上基于EHR的疾病演变过程分析方法存在的两点挑战，提出了一种新颖的基于贝叶斯隐马尔可夫模型（BHMM，Bayesian Hidden Markov Model）[41]-[42][43][44][42][43]的方法来对心衰演变过程进行建模和分析。该模型不限制心衰演变过程只能有一种模式，即本论文提出的模型可以根据真实数据所隐含的信息，学习到各种可能的心衰状态转移路线。此外，为了充分利用EHR数据，该模型的输入包含了EHR中的各种患者信息，如体征、用药、手术及实验室检查等各类信息，使建模结果更具有说服力。

本论文的研究内容主要包括以下几个方面：

* 数据统计与分析。本论文提取了国内某三甲医院心内科的2102名心衰患者的EHR数据作为建模数据集。在正式建模之前，对数据集做了基本的统计分析。首先统计了数据集中各种类型的患者特征（如入院体征、实验室检查、药物干预等）的分布情况。随后选取了三名患者，依次分析他们的入院轨迹，探究心衰演变过程的特点。
* 基于贝叶斯隐马尔可夫模型的心衰演变过程建模与评估。本论文拓展了最新的贝叶斯隐马尔可夫模型，将其应用到心衰演变过程建模的领域。进行一系列实验验证提出模型的可行性和有效性。并提出评估方案，将本论文提出模型与基准模型进行评估对比。
* 心衰患者诊疗数据统计和演变过程分析系统设计与实现。为了便于临床医务人员理解和使用，本论文设计并实现了心衰数据分析系统，系统提供群体统计和个体分析功能，医护人员可以查看感兴趣患者群体的各项统计信息，以及在线查看心衰各个状态的具体含义及特定患者的状态转移路线，为医护人员制定治疗措施和评估干预效果提供参考意见。

### 论文内容安排

本论文一共分为五章。

第一章——绪论。介绍了心衰等慢性病的概念和危害，疾病演变过程分析及建模的概念及研究意义。讨论了基于队列研究和基于EHR的疾病演变过程分析研究的现状，分析其不足之处，并据此说明本论文主要解决的问题及研究内容。

第二章——数据分析。对本论文实验所用的心衰数据集进行分类说明和统计分析。同时为了对心衰演变过程有一个直观的了解，选取了三名心衰患者分析比较他们多次入院的记录。

第三章——基于贝叶斯隐马尔可夫模型的心衰演变过程建模和评估。首先介绍了贝叶斯理论和隐马尔可夫模型的原理，然后阐述本论文提出的贝叶斯隐马尔可夫模型的定义。使用心衰患者EHR数据集对模型进行训练，将其结果和朴素贝叶斯模型进行对比验证提出模型的有效性。

第四章——心衰患者诊疗数据统计和演变过程分析系统设计与实现。基于本论文提出的模型，设计并实现了心衰数据统计及演变过程分析系统，对各项患者特征在群体中的分布情况进行在线统计，同时对特定心衰患者历次入院所处的状态进行推断，帮助医护人员制定合理的治疗计划。

第五章——总结与展望。总结了本论文所做的工作和取得的成果，讨论了现阶段工作的不足之处，并对后续研究工作进行了展望。

# 数据分析

本论文选用心衰这一种典型的慢性心血管疾病来分析其演变过程，采用的数据集是来自国内某三甲医院心内科的2102份心衰患者的EHR数据，包含了患者的体征信息、检查检验信息、用药情况等各类数据。为了充分利用EHR数据，本论文采用了各种类型的综合数据作为患者特征。本章将对这些数据进行统计分析，了解各项患者特征在数据集中的分布情况，并通过几个患者样本多次入院组成的心衰演变轨迹来分析心衰演变过程的特点。

## 基线数据统计

本论文采用的EHR数据集包含2102份心衰患者的13546次入院记录，记录的患者信息包括基本信息、实验室检查结果、用药信息、手术信息、诊断信息、心功能分级及超声心动图检查结果这七大类。本节将分别对这七类数据进行简要的统计分析，以展示各类数据在患者中的分布情况。需要说明的是，对于一些有多次检查结果的特征项（如各项体征和实验室检查结果），本节选取第一次的结果作为患者的基线数据，代表患者在患病初期时的状态。此外，各个统计表格中加粗显示的特征项代表出现频率较高或指标明显异常的特征项。

### 基本信息统计

本论文统计的基本信息包括患者群体的平均入院次数，性别、年龄等人口学特征以及血压、脉搏等体征指标，如表2.1所示：

表2.1 基本信息统计

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **特征** | **平均值/个数** | **标准差** | **单位** | **发生人数** |
| 入院次数 | 6.44 | 4.18 | 次 | 2102 |
| 院内死亡 | 244 | - | 人 | 244 |
| 年龄 | 62.25 | 14.36 | 岁 | 2102 |
| 性别 | 1253/849 | - | M/F | 2102 |
| 舒张压 | 77.97 | 14.32 | mmHg | 1546 |
| 收缩压 | 134.88 | 22.34 | mmHg | 1546 |
| 脉搏 | 79.06 | 14.36 | bpm | 1546 |
| 入院BMI | 25.06 | 3.97 | km/m^2 | 1513 |

为了全面深入地研究心衰演变的过程，在数据预处理阶段筛选掉了入院次数小于3次的患者，最终数据集中所有患者的平均入院次数为6.44次。患者平均年龄为62.25岁，说明老年患者居多，但样本集涵盖了各个年龄段的患者，因此标准差较大。至于血压、BMI等体征指标，平均值处于正常偏高的水平。

### 实验室检查结果统计

本论文一共从EHR数据集中提取到了43项实验室检查指标，其分布情况如表2.2所示：

表2.2 实验室检查结果统计

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **检查指标** | **平均值** | **标准差** | **单位** | **检验人次** |
| **脑利钠肽** | 4353.77 | 8079.92 | pg/ml | 1867 |
| 钾 | 4.02 | 0.52 | mmol/L | 2021 |
| 钠 | 140.57 | 4.16 | mmol/L | 2016 |
| 钙 | 2.22 | 0.12 | mmol/L | 1979 |
| 镁 | 0.868 | 0.087 | mmol/L | 1968 |
| γ-谷氨酰基转移酶 | 43.32 | 52.43 | U/L | 1978 |
| **肌酐** | 131.73 | 180.34 | umol/L | 2028 |
| 尿素 | 7.87 | 3.921 | mmol/L | 2000 |
| 血清尿酸 | 359.7 | 122.2 | umol/L | 1999 |
| 无机磷 | 1.05 | 0.216 | mmol/L | 1961 |
| 葡萄糖 | 6.57 | 2.89 | mmol/L | 2017 |
| 肌钙蛋白T | 0.62 | 3.21 | ng/ml | 1073 |
| **C-反应蛋白测定** | 1.93 | 4.56 | mg/dL | 1048 |
| 低密度脂蛋白胆固醇 | 2.60 | 1.01 | mmol/L | 1644 |
| 高密度脂蛋白胆固醇 | 1.11 | 0.33 | mmol/L | 1639 |
| 丙氨酸氨基转移酶 | 26.25 | 57.45 | U/L | 1984 |
| 天冬氨酸氨基转移酶 | 29.45 | 101.0 | U/L | 1995 |
| 碱性磷酸酶 | 65.87 | 26.84 | U/L | 1976 |
| 肌酸激酶 | 120.20 | 347.29 | umol/L | 1984 |
| 乳酸脱氢酶 | 198.6 | 154.3 | U/L | 1981 |
| **血浆D-二聚体测定** | 0.94 | 1.82 | ug/ml | 1843 |
| 总胆固醇 | 4.47 | 1.38 | mmol/L | 1787 |
| 甘油三脂 | 1.67 | 1.22 | mmol/L | 1786 |
| 总蛋白 | 67.12 | 5.902 | g/L | 1983 |
| 血清白蛋白 | 38.54 | 5.27 | g/L | 1999 |
| 总胆红素 | 15.92 | 13.10 | Umol/L | 1984 |
| 直接胆红素 | 5.58 | 9.184 | umol/L | 1984 |
| 血浆纤维蛋白原测定 | 27.29 | 85.23 | g/L | 1872 |
| 红细胞计数 | 4.34 | 0.66 | 10^12/L | 1977 |
| 血红蛋白测定 | 132.4 | 19.97 | g/L | 1977 |
| 表2.2（续） | | | | |
| **检查指标** | **平均值** | **标准差** | **单位** | **检验人次** |
| 红细胞比积测定 | 1.92 | 24.29 | L/L | 1989 |
| 血小板计数 | 193.0 | 61.5 | 10^9/L | 1977 |
| 红细胞体积分布宽度 | 13.11 | 2.78 | 10^9/L | 1977 |
| 白细胞计数 | 6.68 | 2.76 | 10^9/L | 1616 |
| 中性粒细胞 | 0.644 | 0.74 | % | 1977 |
| 淋巴细胞 | 0.28 | 0.35 | % | 1977 |
| 单核细胞 | 0.07 | 0.05 | % | 1977 |
| 嗜酸性粒细胞 | 0.02 | 0.02 | % | 1977 |
| 嗜碱性粒细胞 | 0.01 | 0.09 | % | 1976 |
| 平均血小板体积 | 10.89 | 1.01 | fL | 1954 |
| 凝血酶时间测定 | 34.22 | 35.98 | s | 1987 |
| 血浆凝血酶原时间测定 | 13.85 | 7.18 | s | 1890 |
| 血浆凝血酶原活动度测定 | 84.82 | 30.40 | % | 1773 |

根据表2.2，有几项指标的平均值明显异常，依次为：

* 脑利钠肽（Brain Natriuretic Peptide, BNP）：是一种心力衰竭标志物，正常值是100pg/ml左右。50岁以下，BNP大于450pg/ml提示异常；50到75岁之间，BNP大于900pg/ml提示异常；大于75岁，BNP大于1800pg/ml，心衰可能性大[45]。由于本论文实验对象为心衰患者，且根据表2.1患者年龄偏大，因此BNP平均值远高于正常值。
* 肌酐：是检验肾脏功能是否出现异常的重要指标之一，一般正常值为：男性60～110 umol/L，女性45～90 umol/L。当急、慢性肾小球肾炎等使肾小球滤过功能减退时，血肌酐升高[46]，由表2.2可知实验患者群体肌酐平均值超出正常范围，提示一部分患者肾功能不全。
* C-反应蛋白（C-reactive protein，CRP）：是血浆在组织受到损伤或人体受到感染时一些快速上升的蛋白质。CRP是一种高度灵敏的炎症标志物，直接参与了炎症与动脉粥样硬化等心血管疾病，是心血管疾病强有力的风险因子。

正常范围值为：800-8000ug/L[47]。

* 血浆D-二聚体：正常值小于0.2ug/L，其值升高的情况多见于继发性纤维蛋白溶解功能亢进，如高凝状态、弥散性血管内凝血、肾脏疾病等。心肌梗死、肺栓塞、静脉血栓形成、手术、肿瘤、弥漫性血管内凝血、感染及组织坏死等也可导致D-二聚体升高[48]。根据表2.2，本论文收集的患者群体的血浆D-二聚体明显高于正常值。

### 用药信息统计

本论文根据药物的原理和用途对患者服用的药物进行归类，如常见的华法林、比伐卢定都归类到抗凝药物，最终得到如表2.3所示的11种类型的药品统计信息。

表2.3 患者服用药品信息统计

|  |  |
| --- | --- |
| **药品类别** | **使用次数** |
| 抗凝药 | 3449/13546 |
| **Beta受体阻滞剂** | 6396/13546 |
| ARB | 3059/13546 |
| ACEI | 1697/13546 |
| 利尿剂 | 4875/13546 |
| **抗血小板** | 6024/13546 |
| **他汀** | 5802/13546 |
| 钙通道阻滞剂 | 5010/13546 |
| **抗酸剂** | 6465/13546 |
| 强心药 | 1989/13546 |
| 扩血管药 | 5275/13546 |

各类药品中使用频率最高的是抗酸剂、Beta受体阻滞剂和抗血小板。这三种药品的功效分别为：

* 抗酸剂：是一类用来中和酸的药物，可以起到降低胃内酸度从而减弱胃液消化作用、降低胃蛋白酶活性的作用，常用于治疗胃溃疡、胃酸分泌过多症等胃病[49]。
* Beta受体阻滞剂：是一类用来治疗心律不齐、防止心脏病二次发作的药物，也可以用来治疗合并心衰、心肌梗塞、心绞痛或心房颤动的高血压患者。是治疗心血管疾病的最常用的药物之一[50]。
* 抗血小板：用来抑制血小板的环氧化酶生长，从而抑制血小板的黏附和聚集，在血栓形成、动脉粥样硬化等过程中起着重要作用，常用来治疗急性冠脉综合征、缺血性脑卒中等心脑血管疾病[51]。
* 他汀：目前最有效的降脂药物，可以抑制肝脏合成胆固醇，最终降低总胆固醇水平。临床上主要用于降低血脂、稳定粥样斑块，是防治冠心病等心脑血管疾病的基础药物[52]。

### 手术信息统计

本论文统计了7种常见的心血管疾病手术的实施情况，如表2.4所示：

表2.4 手术信息统计

|  |  |
| --- | --- |
| **手术名称** | **发生次数** |
| **再血管化手术** | **1223/13546** |
| 瓣膜置换 | 212/13546 |
| 心脏射频消融 | 87/13546 |
| 植入心脏起搏器 | 203/13546 |
| 植入心脏除颤器 | 27/13546 |
| **冠脉造影** | **1975/13546** |
| 冠状动脉旁路移植术 | 111/13546 |

从表2.4可以看出，进行过冠脉造影和再血管化手术的入院次数远远大于进行其他手术的入院次数。其中冠脉造影是用来诊断冠状动脉粥样硬化性心脏病（冠心病）的一种常用方法，被认为是诊断冠心病的“金标准”[53]。而再血管化手术的目的是使缺血心肌恢复供血，解除由冠状动脉狭窄所引起的心肌缺血，防止再梗死的发生，一般有冠状动脉搭桥（CABG）和经皮冠状动脉介入（PCI）这两种方式[54]。

### 合并症统计

本论文对患者群体中高血压等14种主要的心衰合并症的诊断情况进行了统计，如表2.5所示：

表2.5 心衰主要合并症诊断信息统计

|  |  |
| --- | --- |
| **合并症** | **发生次数** |
| 高脂血症 | 3861/13546 |
| 心包疾病 | 483/13546 |
| 脑梗/脑出血 | 2473/13546 |
| 糖尿病 | 4887/13546 |
| 心律失常 | 4032/13546 |
| 贫血 | 2705/13546 |
| 瓣膜病 | 1355/13546 |
| **肾功能不全** | **5438/13546** |
| 甲亢或甲减 | 1564/13546 |
| 心肌梗死 | 2593/13546 |
| 先天性心脏病 | 128/13546 |
| **高血压** | **9217/13546** |
| 心肌病 | 3753/13546 |
| **冠心病** | **8574/13546** |

由表2.5可知，高血压、冠心病和肾功能不全这三种合并症的出现频率最高，这也和表2.3中抗酸剂等药物的大量使用和表2.4中冠脉造影及再血管化手术的多次进行对应起来。

### 心功能分级统计

心功能分级是临床上常用的一种评估心功能受损程度的方法，是指将心脏疾病患者按心功能状态进行分级，从而从大体上反映患者病情的严重程度。本论文的EHR数据集中心功能分级的方法采用的是纽约心脏病协会（NYHA）提出的分级标准，具体如下所示[26]：

* I级：患者有心脏病，但体力活动不受限制，一般体力活动不会引起心功能不全症状（如疲劳、心悸、气急或心绞痛等）。
* II级：患者有心脏病，体力活动轻度受限。休息时无症状，一般体力活动即可引起心功能不全症状。
* Ⅲ级：患者有心脏病，体力活动明显受限。休息时无症状，但小于一般体力活动的轻度活动即可引功能不全症状。
* Ⅳ级：患者有心脏病，体力活动严重受限，休息时也有心功能不全或心绞痛症状。

表2.6 心功能状态统计

|  |  |
| --- | --- |
| **心功能状态** | **发生次数** |
| 心功能Ⅰ级 | 654/13546 |
| 心功能Ⅱ级 | 5604/13546 |
| 心功能Ⅲ级 | 870/13546 |
| 心功能Ⅳ级 | 333/13546 |

由表2.6可知，在本论文的EHR数据集所包含的所有入院诊断中，患者心功能状态被诊断为II级的次数最多，即在大部分入院中，患者处于活动轻度受限的状态。

### 超声心动图统计

超声心动图是一种利用超声脉冲测距原理来检查心脏和大血管的解剖结构及功能状态的技术[55]。本研究选取了超声心动图检查结果中的一项重要指标——左心室射血分数（Left Ventricular Ejection Fraction, LVEF）来进行统计。LVEF指左心室的每搏输出量与心室舒张末期容积的比值, 反映心室的射血能力，是判断心衰类型的重要特征之一。正常情况下LVEF≥50%，小于此值即为心功能不全[56]。本论文选用的心衰患者数据集的LVEF统计结果为：

表2.7 射血分数统计

|  |  |  |  |
| --- | --- | --- | --- |
| **项目** | **平均值** | **标准差** | **发生次数** |
| 射血分数（LVEF） | 51.81 | 12.16 | 611 |

根据表2.7，患者群体射血分数平均值为51.81%，在正常范围内。但要指出的是，并非所有心衰患者的射血分数都会偏低。临床上按照射血分数是否正常将心衰分为射血分数保留型心衰和射血分数降低型心衰，前者就是指患者有心衰的体征和症状，但射血分数正常。

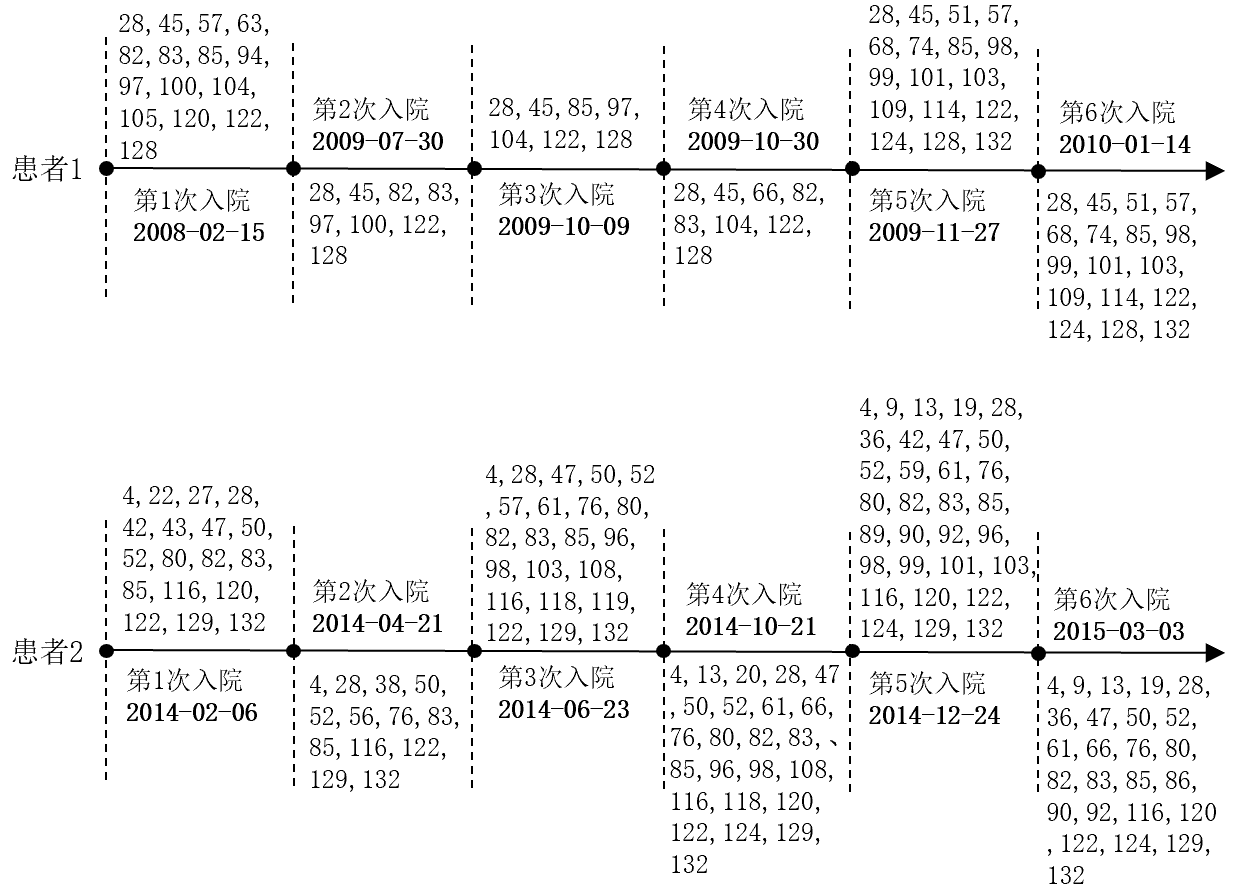
## 案例分析

为了对心衰演变过程有一个直观的了解，本节将从数据集中选取三名心衰患者作为案例进行分析。为了方便表示，首先将EHR中包含的各个临床观察项进行编号得到如图2.1所示的132项特征标签。



图2.1 各类临床观察项标签

需要说明的是，对于用数值表示的特征项，如年龄等体征项及葡萄糖等实验室检查结果，本论文按照其数值范围对应的临床含义将其划分为了三个等级：L（偏低）、N（正常）和H（偏高），将其以后缀的形式标记在图2.1中相关特征项名称的末尾。部分指标如脑利钠肽只存在N（正常）和H（偏高）两个分类。此外，由于本节只关注患者异常的特征项，因此省略了正常项。



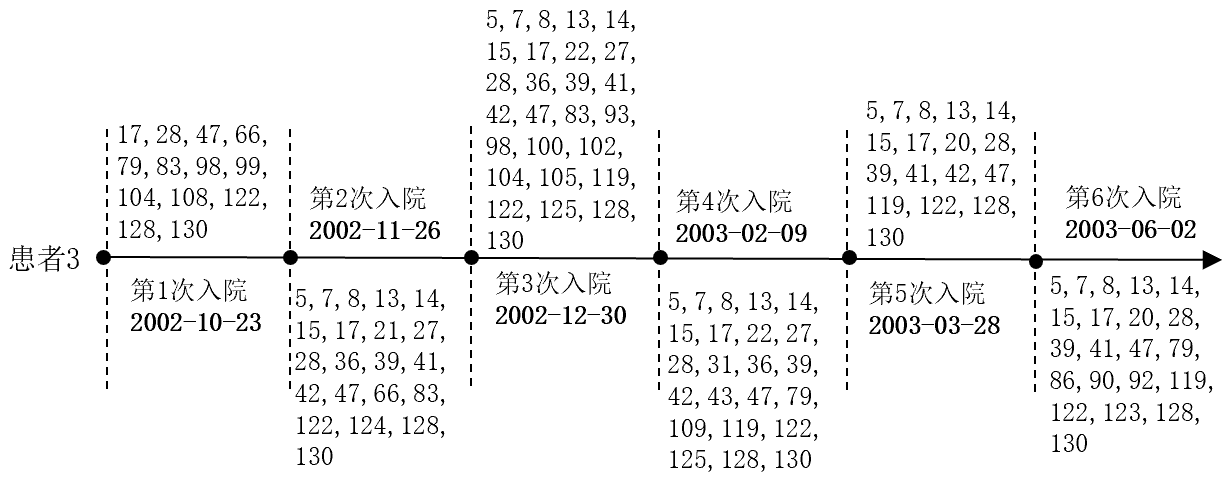


图2.2 三名患者的心衰演变轨迹

图2.2为从EHR数据集中选取的三名入院次数均为6次的心衰患者入院轨迹，每一次入院的数字组合即为与图2.1对应的特征项的编号组合。下面对每名患者历次入院期间观察到的特征项进行概括总结：

* 患者1于2008年2月第一次入院就诊，出现的异常主要是几个实验室检查指标，如甘油三脂偏高、淋巴细胞偏高等，并使用了利尿剂，但没有诊断出任何合并症。一年半后于2009年7月第二次入院，患者的异常特征项较第一次没有发生明显变化，并且接下来的两次入院情况也类似。第五次入院时，异常指标项略有增多，新增了如总蛋白偏高、无机磷偏低等异常项，但依然没有其他合并症出现，也没有使用新的药品、手术干预。最后一次入院与第五次类似。
* 患者2第一次入院时间是在2014年，相比患者1，患者2初次入院患有糖尿病这一合并症，其他异常项与患者1极为类似，主要是一些蛋白、脂等实验室检查指标异常，两个月后第二次入院，情况与第一次类似。从第三次入院开始，出现了脑利钠肽偏高、C-反应蛋白偏高以及血浆D-二聚体偏高等这些重要标志物的异常。第四次入院出现了高血压，第五次入院出现了肾功能不全，并新增了ARB抑制剂干预。第六次入院基本维持在第五次入院的状态。
* 患者3的入院时间最早，首次入院在2002年10月，之后基本每隔1-2个月就再次入院。首次入院时患有肺病且脑利钠肽偏高，其他实验室检查异常指标与患者1和2首次入院类似。但第二次入院时异常项突然增多，出现了心律失常、瓣膜病、高血压、冠心病等合并症，新增了ARB、β受体阻滞剂及抗血小板等药物干预，后面四次入院基本都与第二次相同。

由以上对这三名患者入院轨迹的概括可知，患者1始终保持在病情较为轻微的阶段，没有其他合并症出现，使用的医疗干预也较少；患者2前两次入院和患者1类似，但之后的入院病情逐渐加重，出现了高血压、肾功能不全等重要合并症；而患者3只有第一次入院时症状较为轻微，第二次入院时病情就急剧恶化，出现了心律失常、瓣膜病等合并症，之后几次入院都处于比较严重的状态。此外，三名患者历次入院之间的间隔也不相等。

因此，心衰演变的轨迹是因患者而异的，不同患者之间差异较大，而不是遵循某一种固定的演变模式。不仅演变的趋势不同，速度也大不相同，这就使得对个体患者做精准的心衰演变过程分析变得复杂和困难。因此，需要一种方法能够挖掘出隐藏在复杂多变的诊疗数据背后的心衰状态信息，即利用各项患者特征在历次入院记录中的共现信息抽取出一组心衰状态集合，并确定心衰状态之间的转移关系，由此可以用心衰状态的转移路线来描述心衰演变轨迹，得到心衰演变过程的低维表达，以便针对患者个体做出精准的心衰演变过程分析。

## 本章小结

本章首先对本论文采用的心衰数据集进行了基本的统计分析，分别对基本信息、实验室检查信息、用药信息、手术信息、诊断信息、心功能分级及超声心动图信息这七大类型的数据进行分类统计，得出每个患者特征在整个心衰数据集中的整体分布情况，并对其中的高频出现的特征项进行补充说明。

为了初步探究心衰演变过程，本章还挑选了三名心衰患者作为案例进行分析。具体来说，给出他们历次入院期间观察到的异常项，以此分析病情演变的过程和特点。通过分析发现心衰演变轨迹并不是单一固定的，会因患者而异。因此需要探究一种方法来发现心衰的状态信息以及各状态之间转移的路线。

# 基于贝叶斯隐马尔可夫模型的心衰演变过程建模

本章提出了一种新颖的概率生成模型来对患者的心衰演变过程进行建模，该模型将心衰状态看成患者特征项的概率分布，对EHR的生成过程进行模拟，再通过参数估计得到各个心衰状态。并且使用第二章中的EHR数据集进行了模型的训练和相关实验，最后对模型的有效性进行评估。

## 基础知识

在详细介绍本论文提出的模型之前，先介绍一下该模型应用到的基础知识：贝叶斯网络和隐马尔可夫模型。

### 贝叶斯网络

贝叶斯网络(BN, Bayesian Network) 于1985年由Judea Pearl[57]首先提出，又称信念网络(Belief Network)或有向无环图(DAG, Directed Acyclic Graphical)模型，是一种概率图模型。它是一种模拟人类推理过程中因果关系的不确定性处理模型，其网络拓扑结构是一个有向无环图，表示一组随机变量及其相互之间的条件依赖关系[58],[59]。

贝叶斯网络使用有向无环图中的节点代表随机变量，它们是可以观察到的变量，或潜在变量、未知参数等。连接两个节点的有向边代表这两个随机变量具有因果关系（或非条件独立），其中有向边起点处的节点称为父节点，末尾处的节点称为子节点；而两个节点之间若没有有向边相连就称这两个随机变量彼此间为条件独立[60]-[61][62]。例如，假设节点E直接影响到节点H，即E→H，则用从E指向H的箭头建立节点E到节点H的有向边(E,H)，权值(即连接强度)用条件概率P(H|E)来表示，如下图所示：

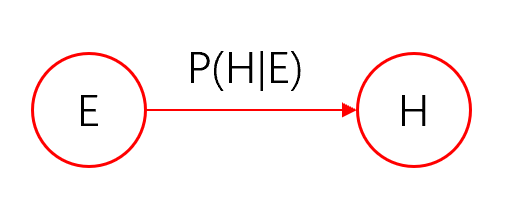


图3.1 贝叶斯网络中具有因果关系的两个节点

贝叶斯网络的数学定义[63]-[64][65][66][67][68]如下：

令 G= (I,E) 表示一个有向无环图，其中 I 代表图中所有的节点的集合，而 E 代表所有有向边的集合，令X= Xi (i∈I)为其有向无环图中的某一节点 i 所代表的随机变量，若节点 X 的联合概率分布可以表示成：

(3.1)

则称 X 为相对于有向无环图 G 的贝叶斯网络，其中 pa(i)表示节点 i 之「因」, 或称pa(i)是i的parents（父母）。

对任意的随机变量，其联合分布可由各自的局部条件概率分布相乘而得出：

一个简单的贝叶斯网络如下图所示：

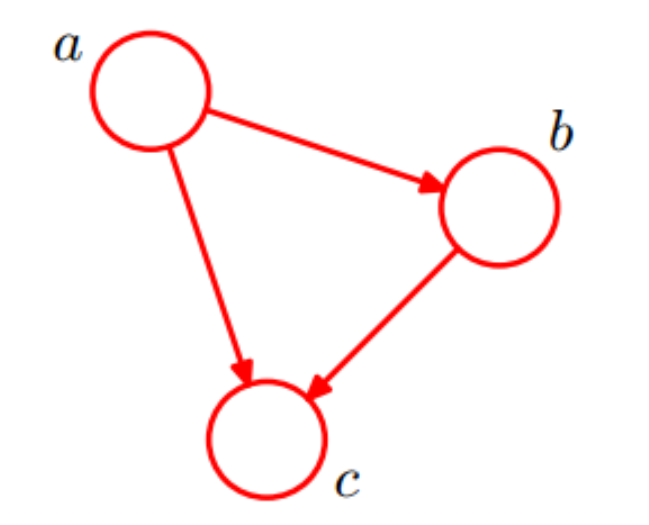


图3.2 一个简单的贝叶斯网络

因为a影响b，a和b影响c，所以a,b,c的联合分布概率为：

### 隐马尔可夫模型

隐马尔可夫模型（Hidden Markov Model, HMM）由Leonard E. Baum等人[69]-[70][71]][67][于20世纪提出,是一种关于时序的概率模型，描述由一个隐藏的马尔可夫链随机生成不可观测的状态序列(state sequence)，再由各个状态生成一个观测而产生观测序列（observation sequence）的过程。HMM作为一种经典的机器学习模型，它在语音识别、故障诊断及模式识别等领域得到了广泛的应用[72],[73]。

首先介绍一下马尔可夫链（Markov Chain, MC）。马尔可夫链也叫马尔可夫过程，是在贝叶斯网络中概率图退化成线性链时的特殊情况[74],[75]，如下图所示：



图3.3 马尔可夫链

马尔可夫链描述了状态空间中从一个状态到另一个状态的转换的随机过程。这个过程需要满足马尔可夫性质：当前状态只和上一个状态有关，而和之前的任何其他的状态无关[76],[77]。即在图3.3中，的分布状态只和有关，和其他变量条件独立。

HMM是在马尔可夫链的基础上发展而来的，可以看作一个双重的随机过程。在基本的马尔可夫随机过程中，状态序列就是可以直接观察到的观测序列。而HMM的状态序列是不可直接观测到的，观测值是由“隐藏的”状态随机产生的[78]-[79][80]。HMM的结构如图3.4所示：

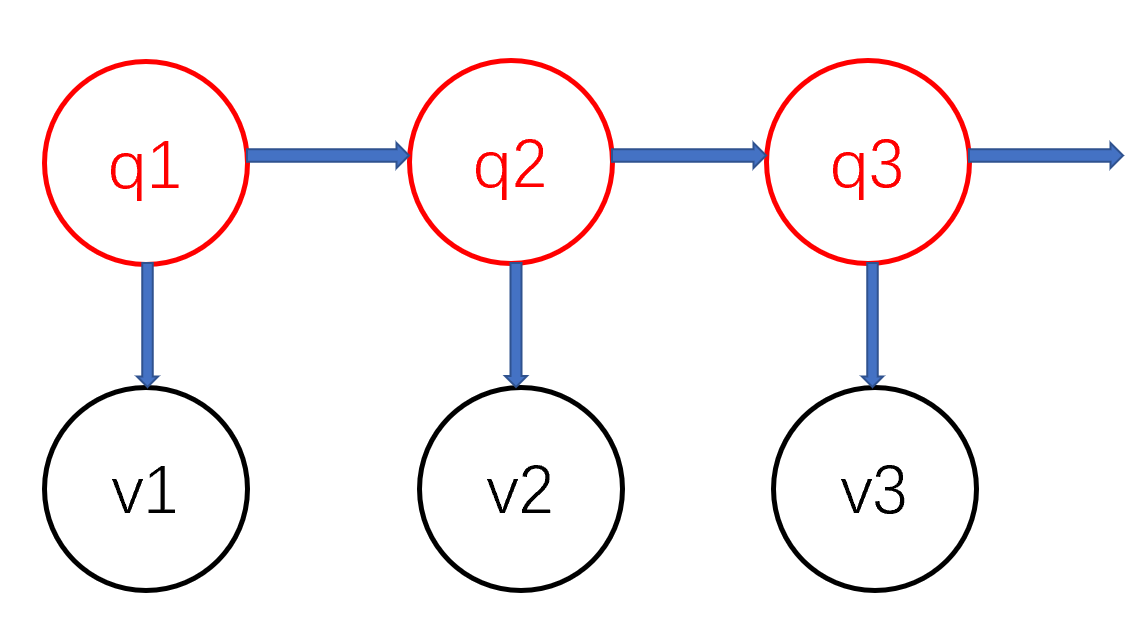


图3.4 HMM结构示意图

设Q = {,…,}是所有可能的状态的集合，V = {,…,}是所有可能的观测的集合，其中N是所有可能的状态数，M是所有可能的观测数。设I = {,…,}是长度为T的状态序列，O = {,…,}是对应的观测序列。HMM由初始状态概率分布**π**、状态转移概率分布A，观测概率分布B这三个参数确定，记为λ= (π, A, B)，下面分别对每个参数进行说明[78]：

* 初始状态概率分布**π={}，**其中

(3.4)

代表时刻t=1时处于状态的概率。

* 状态转移矩阵A={}，其中

表示在时刻t处于状态的条件下，在时刻t+1转移到的概率。

* 观测概率矩阵B={}，其中

表示在时刻t处于状态的条件下生成观测的概率。

HMM的三个基本问题[78]为：

1. 概率计算问题：给定模型λ= (π, A, B)和观测序列O = {,…,}，计算在模型λ下出现该观测序列的概率P(O|λ)，是最基本的问题。
2. 学习问题：已知观测序列O = {,…,}，估计HMM的参数λ= (π, A, B)，使得在该模型参数下观测序列O出现的概率P(O|λ)最大。这个问题也是本论文提出的模型要解决的问题，实际上是通过已知的观测序列训练出一个HMM。
3. 预测问题：也称为解码问题，已知HMM参数λ= (π, A, B)和观测序列O = {,…,}，求一个状态序列I = {,…,}使得条件概率P(I|O)最大。即产生给定观测序列对应的最有可能的状态序列。

## 模型构建

本节会详细地介绍模型的细节，分为两部分：

* 心衰演变过程建模问题的定义
* 模型的概率图表示和算法实现

### 心衰演变过程建模问题的定义

在本论文中，用集合来表示|D|个心衰患者的EHR记录，用集合代表在中出现的所有临床观察项，包括患者的人口统计学、体征、实验室检查结果、用药及手术等信息。每个EHR ()对应一个特定的患者，由一组患者入院记录组成，即，其中包含一系列临床观察项，它们描述了患者在第t次入院就诊时的心衰状态，即，其中，是患者d在整个治疗轨迹中总的入院次数。

需要注意的是，多次入院之间存在着某种顺序依赖性。具体来说患者每次入院的EHR，描述了患者在该次入院时所处的心衰状态，这是一个隐含的抽象变量，可以表示为关于一组临床观察项的多项分布。实际上，心衰状态z就是上一节介绍的隐马尔可夫模型（HMM）中的“隐藏”的变量。也就是说，将患者的多次入院看作一个隐马尔可夫过程，每次入院对应一个隐藏的状态，而每次入院的EHR中记录的一组临床观察项则是该隐藏状态下产生的可观测项，且多次入院对应的心衰状态序列之间存在一个状态转移概率矩阵。

利用一组EHR集合中某些临床观察项经常同时出现的信息，无监督的学习方法可以自动发现代表某个心衰状态的一组高度相关的临床观察项。而一旦一组高度相关的临床观察项从EHR中被提取出来，医学研究者就可以给它们分配有意义的标签用于进一步的应用中，例如本论文中的心衰演变过程分析。本论文要解决的问题定义如下：

**问题** 给定一个心衰EHR集合，其中每个EHR 包含一系列入院记录，每次入院 具体表示为在该次入院期间产生的一组临床观察项，包括患者的症状、体征及用药等干预措施，这些信息反映了患者在该次入院所处的隐藏的心衰状态。假设所有的患者样本共享K个隐藏的心衰状态。则心衰演变过程建模的目的就是从每个患者治疗轨迹的EHR中识别出每次入院所处的心衰状态以及多次入院之间心衰状态转移的路线。

### 心衰演变过程模型的提出

针对上一小节中提出的问题，本节提出了一种基于贝叶斯隐马尔可夫模型的概率生成模型用于心衰演变过程建模（Heart Failure Progression Modeling, HFPM），记为BHMM-HFPM，其概率图如图3.5所示，相关变量的含义如表3.1所示：

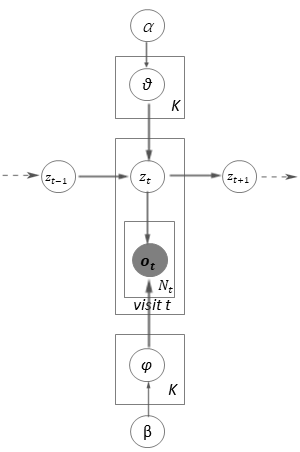


图3.5 基于贝叶斯隐马尔可夫模型的心衰演变过程模型（BHMM- HFPM）

表3.1 本研究中使用的变量

|  |  |
| --- | --- |
|  | 狄利克雷分布（Dirichlet）的先验参数 |
| *K* | 心衰状态的数目 |
| *O* | 临床观察项的数目 |
|  | 一个患者的治疗轨迹 |
|  | 患者的第*t*次入院 |
|  | 第*t*次入院期间产生的临床观察项的数目 |
| 表3.1（续） | |
|  | 心衰状态转移概率矩阵 |
|  | 第*t*次入院对应的心衰状态 |
|  | 特定于临床状态z的临床表现的多项式分布 |

如图3.5所示，假设在一个患者的治疗轨迹中，第t次入院对应的心衰状态由该次入院产生的临床观察项和上一次入院（即第t-1次）所处的心衰状态决定。这个假设遵循一阶马尔可夫性质，因此可以应用一阶马尔可夫链结构来处理HFPM问题。

具体来说，给定一个患者样本，包含次入院记录，其心衰演变过程可以表示为：

(3.7)

(3.8)

(3.9)

(3.10)

其中公式（3.7）表示心衰状态服从多项分布，指的是当上一次（*t*-1次）入院所处的心衰状态为时，第*t*次入院的心衰状态转移概率分布。在公式（3.8）中，指的是在*d*(t)中记录的第*j*个临床观察项，指的是当心衰状态为时临床观察项的概率分布。此外，、的先验分布分别服从参数为α、β的Dirichlet分布，且本研究采用对称的先验参数。

图3.5所示BHMM-HFPM的生成过程为：

给定一个患者的治疗轨迹*d*和其中的第*t*次入院，参数为α的狄利克雷分布会产生一个心衰状态转移概率分布的先验分布（即），参数为β的狄利克雷分布会产生一个临床观察项的先验分布（即）。本次入院对应的心衰状态由心衰状态转移概率分布和上一次入院对应的心衰状态所确定，最终一个临床观察项o由本次入院所处的心衰状态对应的临床观察项的多项分布产生。因此，BHMM-HFPM模型中观测到的以及隐藏的变量对应的联合分布概率如下所示：

. (3.11)

为了评估出BHMM-HFPM的模型参数，最优方法是使用最大似然估计[80]，使得给定的临床观测项出现的概率最大。然而，很难直接通过数学方法来最大化似然函数，解决方法是使用近似推断，本论文中采用了吉布斯采样（Gibbs Sampling）方法[82]-[83][84][85]。首先，从公式（3.11）中可以推导出每个患者样本每次入院*d*(*t*)的全条件分布（full conditional distribution），如下所示：

(3.12)

其中表示从中排除当前的入院记录*d*(*t*)，代表除了当前入院*d*(*t*)之外的所有入院对应的心衰状态。

接下来采用一个包含两个步骤的吉布斯采样方法来估计式（3.12）中的每个乘积项。第一步，是根据以下公式来采样心衰状态z：

(3.13)

(3.14)

式（3.13）中是指除了当前入院外，所有其他入院诊断中心衰状态为z而其上一次入院的心衰状态为的总次数，分母中的。式（3.14）同理，其中是一个指示函数，即如果为真则等于1，如果为假则等于零。

第二步则是根据以下后验概率得到在*d*(t)中的临床观察项：

(3.15)

其中满足：

其中指除了本次入院外在其他所有心衰状态为z的入院中出现临床观察项的次数。

当采样进行足够多的迭代次数之后，可以通过以下（3.17）和（3.18）所示的后验估计来估计参数和 ：

|  |
| --- |
| (3.17) |
| (3.18) |

算法1为训练BHMM-HFPM的伪代码：

|  |  |
| --- | --- |
| 算法1: 基于贝叶斯隐马尔可夫模型的心衰演变过程建模 | |
| **Input**: an EHR dataset , hyper-parameters , the number of disease states , and the number of iterations | |
| **Initialization step:** | |
|  | Initialize , and |
|  | For each patient sample |
|  | For each visit |
|  | Sample a *d*isease state assignment |
|  | Increment = |
|  | For each clinical manifest |
|  | Increment |
|  | End For |
|  | End For |
|  | End For |
| **Gibbs sampling step:** | |
|  | For each iteration |
|  | For each patient sample |
|  | For each visit |
|  | Decrement = |
|  | For each clinical observational item |
|  | Decrement |
|  | End For |
|  | Sample a disease state assignment using Equation (3.12) |
|  | For each clinical manifest |
|  | Increment |
|  | End For |
|  | Increment = |
|  | End For |
|  | End For |
|  | End For |
| **Output**: Estimate the model parameters with sampling results as follows: | |
|  |  |
|  |  |

模型训练完后，一个重要的任务就是当给定患者样本*d*的一个新的入院记录*d*(*t*)时，根据训练好的模型推断出该次入院对应的心衰状态。具体而言，就是计算每一个心衰状态对应的概率，然后选择概率最大的作为该次入院对应的心衰状态。计算心衰状态的公式如下：

(3.19)

特别地，如果*d*(*t*)是第一次就诊，即没有上一次就诊*d*(*t*-1)，则认为。

## 实验和分析

本节将会使用第二章中的心衰数据集对上面提出的BHMM-HFPM进行训练，评估出模型参数，然后据此进行进一步的实验以深入分析心衰演变过程。实验包括：

* 心衰状态数目的选取
* 心衰状态的发现
* 心衰状态与合并症的关联性
* 心衰状态转移路线

### 实验设置

BHMM-HFPM中的超参数α和β的经验值一般取值为α= 50/K，β= 0.01[86]，起到平滑数据的作用。完成实验的平台是Windows 10操作系统，处理器的型号为Intel® Core™ i5-8265U CPU @ 1.60GHz，运行内存为8GB。模型的训练和所有的实验都是采用Python语言实现。

### 结果与分析

**3.3.2.1心衰状态数目的选取**

**在上一节训练模型的算法1伪代码中，需要输入心衰状态的数目***K***。本论文选取使用最为广泛的方法——基于困惑度（perplexity）[87],[88]的方法来确定***K***的值。具体到本研究中，对于由*D*个患者的EHR数据集，困惑度定义为：**

**其中**指的是在患者*d*的心衰演变轨迹中出现过的临床观察项，分子中的可以表示为：

其中是患者*d*出现的临床观察项的总数目。用不同的*K*重复实验，当困惑度最优时，此时的K即为最佳取值。一般而言，模型困惑度会随着*K*的增大而减小。而另一方面，如果*K*过大则会增加模型的复杂度，出现过拟合的现象。因此需要在模型的困惑度和简洁性之间做一个平衡[89]。

本论文分别设置*K*={1, 3, 5, 7, 10, 13, 15, 20, 25, 30, 35, 40, 45, 50}来训练所提出的BHMM-HFPM，结果如图3.7所示，由图可知，刚开始时模型的困惑度随着心衰状态数目的增加迅速下降，但当*K*大于10时，困惑度的下降速度显著降低。考虑到模型的复杂度，最终选取*K*=10。

**图3.6 困惑度随心衰状态数目的变化**

3.3.2.2 心衰状态的发现

如前所述，本研究提出BHMM-HFPM的目的就是利用EHR数据发现隐含在患者心衰演变轨迹中的心衰状态。为此，实验过程中计算了模型中的参数，然后列出了与每种隐藏的心衰状态高度相关的前40项临床观察项，并按降序排列的方式列在表3.2中，与2.2节相同，表中与实验室检查相关的临床观察项的后缀H、N、L分别代表该项检查指标偏高、正常和偏低，此外加粗显示的为异常的观察项。

表3.2 不同的心衰状态对应的临床观察项分布表



表3.2 （续）



表3.2（续）



表3.2 （续）



从表3.2可以看出，不同的心衰状态下对应的临床观察项的分布概率大不相同。状态1对应的前40项临床观察项大多都是实验室检查指标，而且是正常的，说明状态1下的心衰症状不明显，患者的病情比较轻微。而在状态2下，异常指标主要是“脑利钠肽偏高”、“血浆D-二聚体测定偏高”及各种血细胞减少等，还出现了抗酸剂药物干预，但是没有其他合并症，由2.1节可知，脑利钠肽是心衰的重要标志物，预示着状态2下心衰患者的病情处于中等的状态。状态6下异常观察项主要是冠心病、心源性入院和抗血小板，意味着处于状态6下的心衰患者很有可能因为心脏病发作（冠心病）入院。状态8则对应最多的异常指标，除了“脑利钠肽偏高”外，还有“肾功能不全”、“高血压”和“冠心病”等合并症以及“利尿剂”等干预措施，说明处于状态8的患者病情最为严重。因此，心衰等慢性病的临床表现复杂多变，患者的心衰演变轨迹应该是各种可能的心衰状态的组合，而不是某一种固定的演变模式。

图3.7显示了各个心衰状态出现在不同的患者中的人数。总体来说，出现这10个状态的人数差别不大，即每一种状态对应的临床观察项的组合在心衰患者的病情演变轨迹中都很常见，进一步说明了心衰演变模式的复杂性和多样性。但有了心衰状态与临床观察项的对应关系后，可以用简单的数字来代替繁多的临床观察项去描述患者入院时的状态以及多次入院之间的状况变化（如3.3.2.3所述），实现心衰病情描述及演变过程的低维表达。

图3.7 分配给每种心衰状态的患者数目

**3.3.2.3 心衰状态转移路线**

**模型训练完后得到了各个心衰状态之间的转移概率矩阵，即模型中的参数，如图3.8中的热图所示。对角线上的概率通常远大于非对角线，这表明患者在多次入院之间会倾向于保持相同的状态。但部分状态转移到其他状态的概率也较大，如状态5到状态3，状态7到状态3、5、6等。**

**** 

**图3.8 状态转移矩阵**

得到θ和φ之后，可以根据公式（3.19）推算出患者任意一次入院时所处的状态，基于此统计了心衰状态随入院次数的分布，如图3.9所示。z轴表示在特定入院次数处于特定状态的人数。大多数患者在最初的几次入院都处于状态3、6、1和7，根据表3.2，几个状态对应的异常项比较少，或者是冠心病发作导致心源性入院，处于心衰早期阶段。随着入院次数的增多，处于这几种状态的患者减少，而处于状态2、4、5和9的患者增多，这几种状态都对应不同的异常项，或是多项实验室检查指标异常，或是出现了高血压、心律失常等合并症，预示心衰加重。到最后的入院中处于状态8和10的患者最多，而状态8是最严重的状态，状态10主要对应了几种合并症。说明随着时间的推移，患者病情逐渐加重，部分患者经过治疗维持在状态10，而部分患者病情持续加重到状态8甚至死亡。

**图3.9 心衰状态随入院次数的分布**

**为了直观地说明如何将本论文提出的模型运用于追踪单个患者的心衰演变轨迹上，本小节从实验数据集中选取了两名患者样本来分析他们的心衰演变过程。具体来说，根据公式（3.19）来计算出每个患者每次入院最有可能的状态，结果如图3.10所示。**

**如图，患者1第一次入院时处于状态1，根据表3.2，状态1各项临床观察项基本正常，是病情最轻微的状态，后续病情变重，由于心脏病发作第二次入院就诊，即由状态1转移到状态6，随后又转移到了状态3，并且接下来几次入院都保持在状态3，说明病情被控制得较为稳定；与患者1相反，患者2一开始就患有心血管疾病（心力衰竭、心律失常），且使用了利尿剂、β受体阻滞剂等药物干预，在三次入院后恶化，转移到最严重的状态8，预后不乐观。这两名患者的病情演变过程分析反映了本论文提出的BHMM-HFPM在临床决策支持系统中的潜力，即为医务人员提供合理的预测和干预建议。**

**图3.10 心衰患者个体状态转移路线**

3.3.2.4 不同心衰状态与合并症的相关性

**本小节设计实验来探究合并症在不同心衰状态下的发生情况。具体来说，本实验选取了三种常见的合并症：高血压、冠心病及肾功能不全，分别统计了在所有就诊记录中各个心衰状态下每种合并症出现的次数，结果如图3.11-3.13所示。**

**不同状态下对应的合并症出现频率有明显区别，且不同的合并症与心衰状态之间的关系也互不相同。对于高血压，状态3和状态4对应的高血压出现次数最多，而状态2对应的高血压出现次数最少；而对于冠心病，则是状态3和状态6出现的次数最多，同样也是状态2出现的次数最少；至于肾功能不全，在状态4、8和10中出现的次数远远高于其他状态。这与表3.2中各状态对应的临床观察项的分布概率基本一致，也反映出BHMM-HFPM模型的准确性。**

**图3.11 高血压出现次数与心衰状态的关系**

**图3.12 冠心病出现次数与心衰状态的关系**

**图3.13 肾功能不全出现次数与心衰状态的关系**

## 模型评估

### 评估方案设计

为了定量评估提出模型的准确性，本小节提出了以下的定量评估指标和评估方案：

给定患者样本，取前次的就诊记录来拟合模型，即根据训练好的BHMM-HFPM模型参数和这次的EHR数据推断出第次入院所处的心衰状态，然后用以下公式计算第次入院对应的各临床观察项的出现概率：

然后根据概率大小对临床观察项进行降序排序，取前N个临床观察项，认为会出现在第次入院中。以第次入院时EHR中记录的真实发生的临床观察项作为金标准，计算与相似的程度，记为：

其中代表由模型推导出的前N个临床观察项中，出现在真实的就诊记录中的数目。

本实验依次计算了N=5，10和50时的，并与两个经典的基准模型——潜在狄利克雷分配（Latent Dirichlet Allocation, LDA）[87]和朴素贝叶斯（Naïve Bayes, NB）[90]-[91][92]进行了对比。LDA是一种文档主题生成模型，将患者的每次入院看作一篇文档，而每次入院期间产生的临床观察项则看做是构成文档的词项，通过模拟文档的生成过程识别文档中潜在的主题信息，即每次入院隐含的心衰状态。

### 结果与分析

在上述实验中，输入每个患者样本*d*的前次就诊记录来拟合模型，并使用最后一次（第次）入院中实际发生的临床观察项作为评估的依据。此外，还将本模型与最基本的朴素贝叶斯模型（Naive Bayes Model, NBM）[90]-[91][92]进行了对比。结果如表3.3所示。

结果表明，在N取不同值的情况下，本论文提出的模型评估的准确性都高于基准模型，进一步证明了BHMM的有效性。此外，基于这一特点，还可以提出BHMM-HFPM的进一步应用——临床表现预测，即预测患者治疗轨迹中即将出现的临床观察项。具体来说，给定患者目前的诊疗数据，可以通过公式（3.22）对所有临床观察项出现的概率进行估计，然后根据概率大小进行排序，取前N个作为最有代表性的临床观察项，可以作为患者下次入院身体状况的参考，从而有助于做好早期预防。

表3.3 与NBM比较评估结果的准确性

|  |  |  |  |
| --- | --- | --- | --- |
| 模型类别 | N | | |
| 5 | 10 | 50 |
| BHMM | 0.96 | 0.94 | 0.82 |
| LDA | 0.92 | 0.88 | 0.79 |
| NB | 0.87 | 0.85 | 0.73 |

## 本章小结

本章提出了一种基于贝叶斯隐马尔可夫模型的心衰演变过程建模方法，对患者的心衰演变过程进行深入的分析，该模型是一种概率生成模型，将患者的心衰状态视为模型中的隐藏变量，将包含患者诊疗信息的EHR数据作为观测序列，由此将患者的心衰演变过程视为一个隐马尔可夫过程，通过解决马尔可夫模型的学习问题估计出患者在心衰演变过程中对应的一系列心衰状态及状态之间相互转移的路线。

本章利用第二章中从中国某三甲医院心内科收集的心衰患者的EHR数据进行了一系列实验，实验结果表明本模型发现的心衰状态具有良好的临床意义和可解释性。

最后，为了对模型的有效性进行验证，本章还通过比较提出的模型和基本的朴素贝叶斯模型在预测临床表现上的准确性，结果表明提出模型的准确性高于基准模型，证明了提出模型的准确性和有效性。

# 分析系统设计与实现

本章以第二章和第三章中的工作为基础，设计并实现了心衰患者诊疗数据统计和心衰演变过程分析系统。该系统存储了本论文使用的EHR数据集，具有良好的交互界面，可以展示患者群体的统计信息、在线推断特定患者每次入院对应的状态以及多次入院之间的状态转移路线，从而帮助医务人员深入了解患者的心衰演变过程，制定进一步的诊疗计划。

## 整体框架

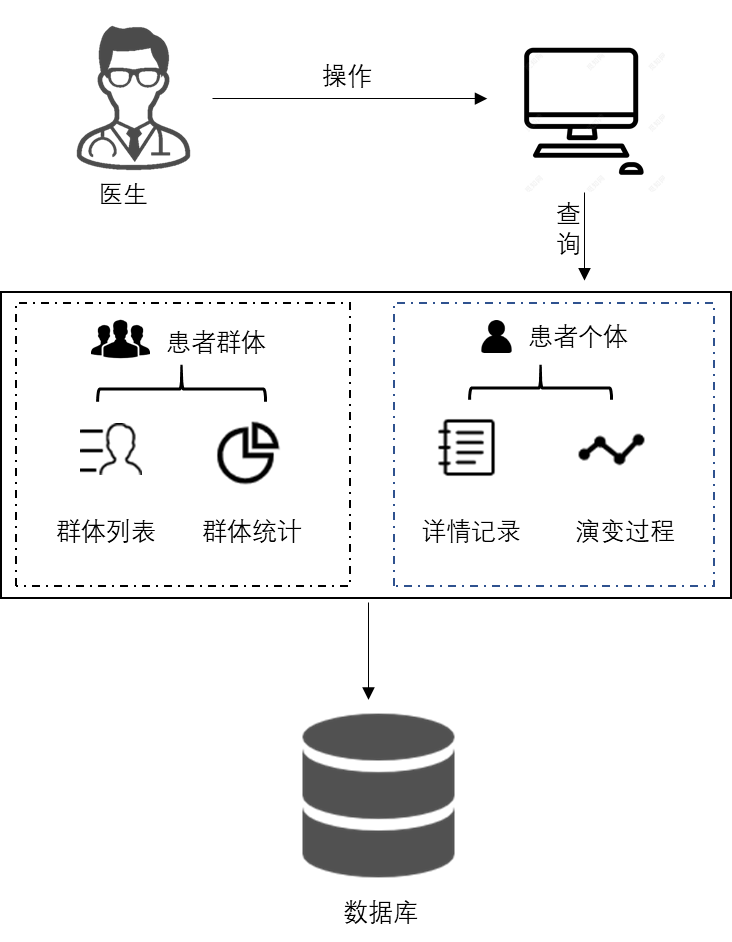


图4.1 系统整体框架示意图

如图4.1所示为心衰患者数据统计与病情演变过程分析系统的总体框架示意图。医务人员在电脑端登录系统，即可通过简单的点击操作使用本系统的各项功能。总体上，该系统包括患者群体统计和个体分析两大模块。具体来说，医务人员可以筛选出自己感兴趣的患者群体，查看该群体中的患者列表以及针对该群体各项特征的在线统计数据；也可以选中或筛选出特定的患者，查看该患者的历次入院记录详细数据，当然更重要的是查看患者每次入院对应本论文提出模型中的哪种心衰状态，进行心衰演变过程分析。

## 功能模块

由上一节系统整体框架示意图可知，系统的功能主要分为患者群体统计和患者个体分析这两大模块，本节将对这两大模块的设计进行说明。

### 群体统计

为了方便医务人员迅速定位感兴趣的研究患者群体，本系统提供了14种筛选器，如图4.2所示。医务人员可以设定筛选条件来筛选出满足条件的患者。

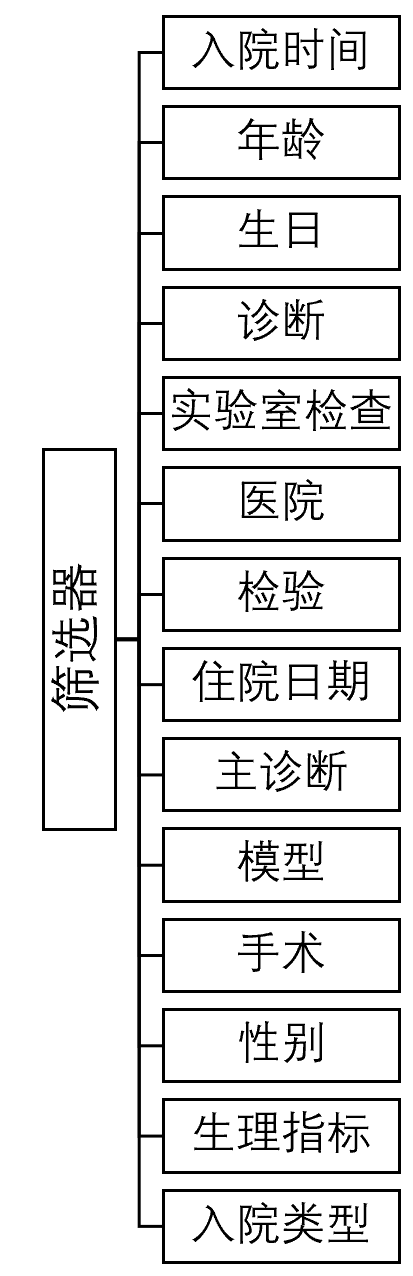


图4.2 患者群体筛选器

筛选出满足条件的患者群体后，医务人员还可以查看该群体的各项统计信息。与本论文2.1节中的基线数据统计类似，本系统中的实时在线统计项包括基本的人口学信息统计、实验室检查结果统计、用药统计以及手术统计等。如此，医务人员可以快速方便地查阅各项特征在感兴趣患者群体中的分布情况。

### 个体详情

医务人员可以在上述患者列表中选择某个特定患者查看该患者的个体详情，患者个体详情包括该患者历次入院产生的EHR数据，即前面章节中提到的患者基本信息、实验室检查、用药记录等，如图4.3所示。更重要的是，系统会利用已部署到服务器中的BHMM-HFPM模型对该患者的EHR数据进行拟合，生成患者每一次入院对应的心衰状态，方便医务人员全面深入地了解患者个体的心衰演变过程。

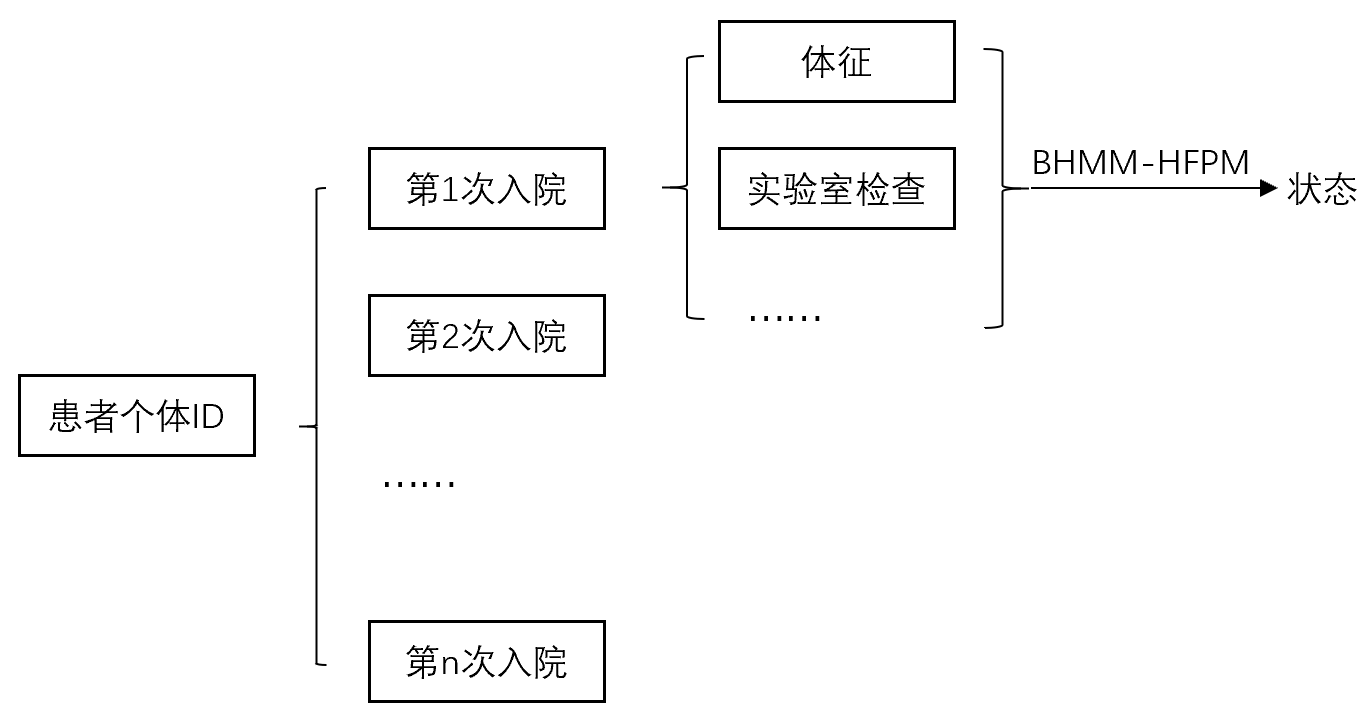


图4.3 个体详情示意图

## 系统实现

本节将展示根据4.1节的整体框架和4.2节的功能模块实现的系统最终效果。本系统的前端技术栈主要使用react+redux+react-router-dom，后端则利用spring boot框架进行搭建，后端提供的接口全部以RESTful API形式提供。根据前面两节可知系统的主要功能页面有：患者筛选、群体列表、群体统计、个体详情。下面依次对这些页面进行展示：

* 患者筛选

如图4.4为系统主界面，医务人员可以点击左侧控制面板的“添加新查询”按钮来进行查询，点击后系统会弹出提示框提示选择查询类型，如果是查询一个确定的患者则选择“个体分析”，否则选择“群体分析”，此处选择群体分析。选定查询类型后会出现一个带有各种筛选条件的弹出框，如图4.5所示，医务人员可以根据需要进行筛选。例如图中选择年龄筛选器，可以输入一个年龄段来筛选出对应的目标患者群体。



图4.4 系统主界面



图4.5 筛选患者

* 群体列表

在上一步中设置好筛选条件后，系统就会筛选出满足相应条件的患者，并以患者列表的形式展示出来，如图4.6所示。每个患者作为列表中的一行，显示患者的基本信息，如患者ID、医院名称等入院信息及姓名、性别等基本的个人信息。

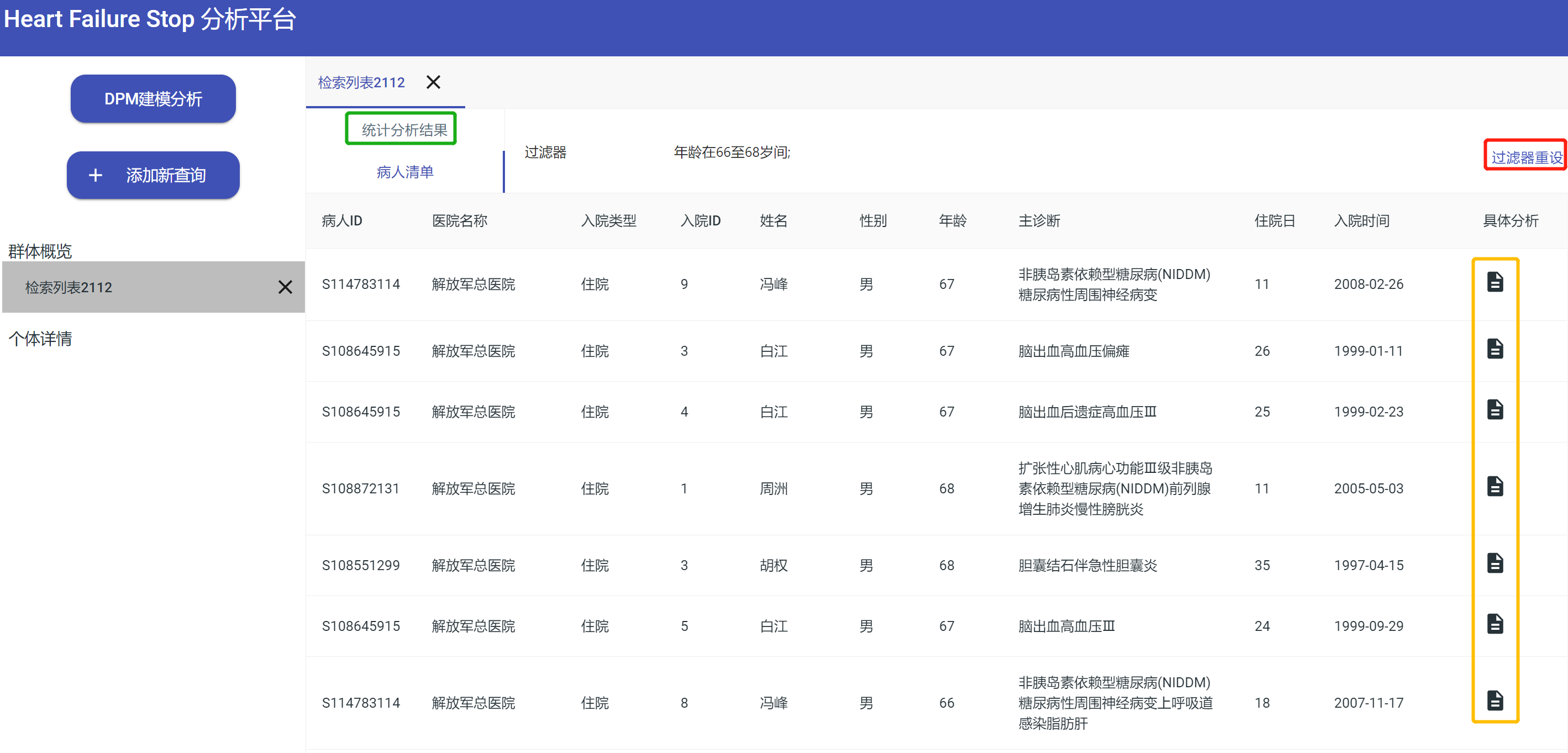


图4.6 符合筛选条件的患者群体列表

此外，该页面还有三个功能：

1. 点击右上角的“过滤器重设”按钮（ 所示）可以重新设置或添加新的筛选条件。
2. 点击上方“统计分析结果”（ 所示），将会跳转到群体统计界面，具体在下一条“群体统计”中论述。
3. 在列表中选择一个特定的患者（ 所示），即可跳转到该患者的详情页面，具体在之后的“个体详情”中论述。

* 群体统计

如前所述，在筛选出的患者列表页面点击“统计分析结果”就会进入到该患者群体的统计分析页面，该页面以丰富的图表展示患者群体的各项统计信息。具体包括：该患者群体的性别、年龄、出院诊断和住诊断统计，如图4.7所示；各项实验室检查指标的统计，如图4.8所示，注意在统计实验室检查指标时，以各患者第一次入院的数据进行统计。最后，还统计了该患者群体的手术和用药分布情况，如图4.9所示。

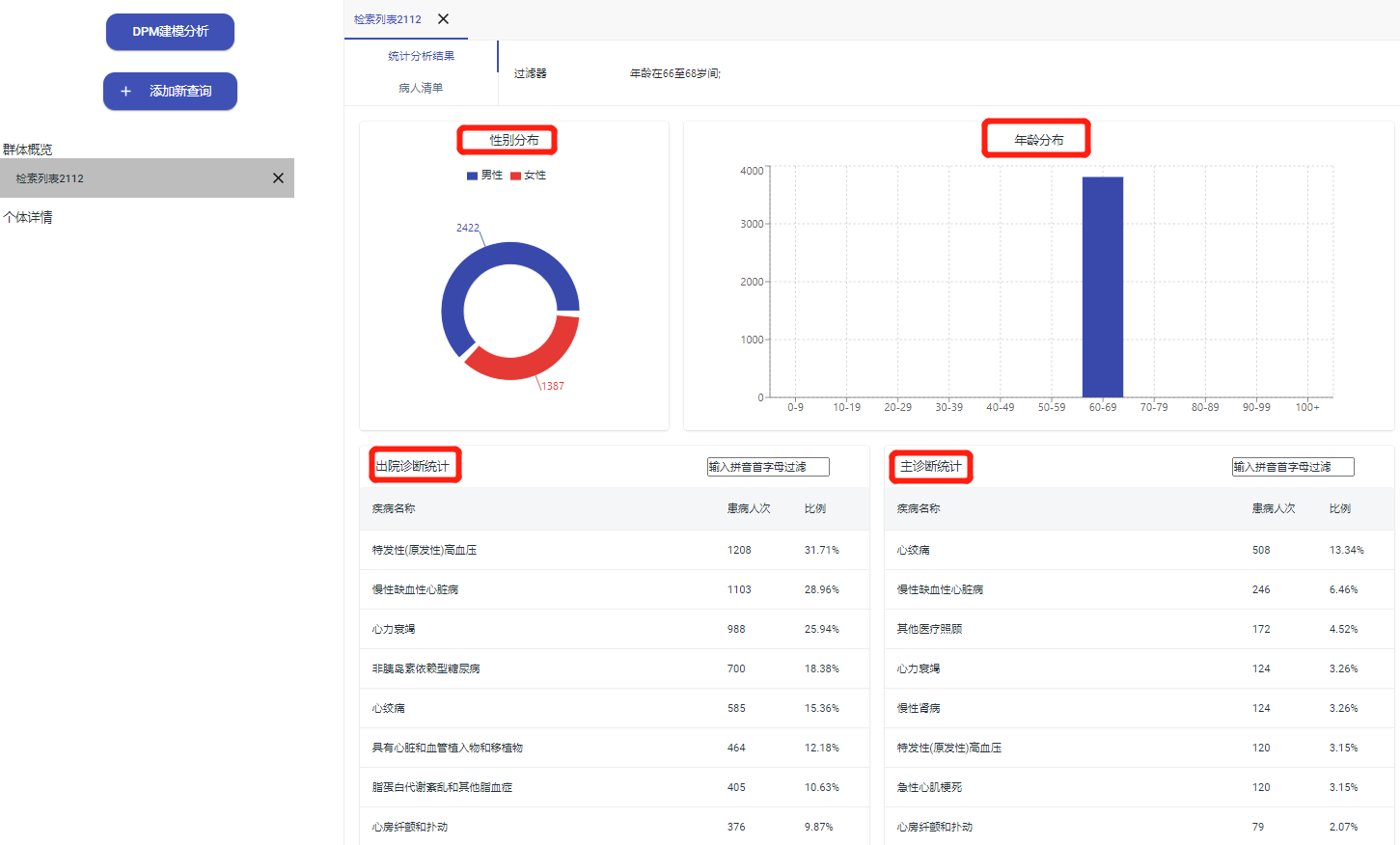


图4.7 患者群体的年龄、性别、诊断统计

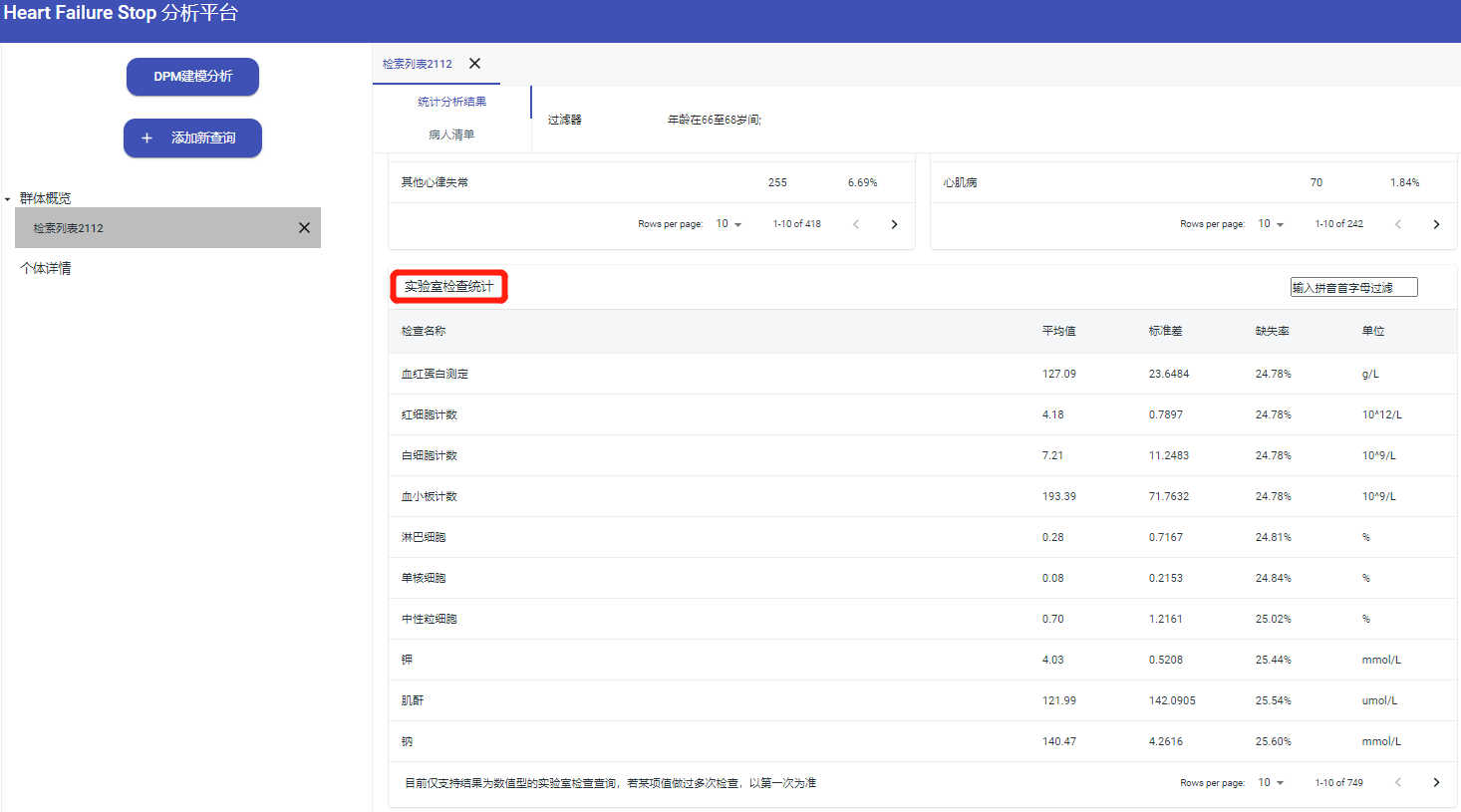


图4.8 患者群体的实验室检查结果统计

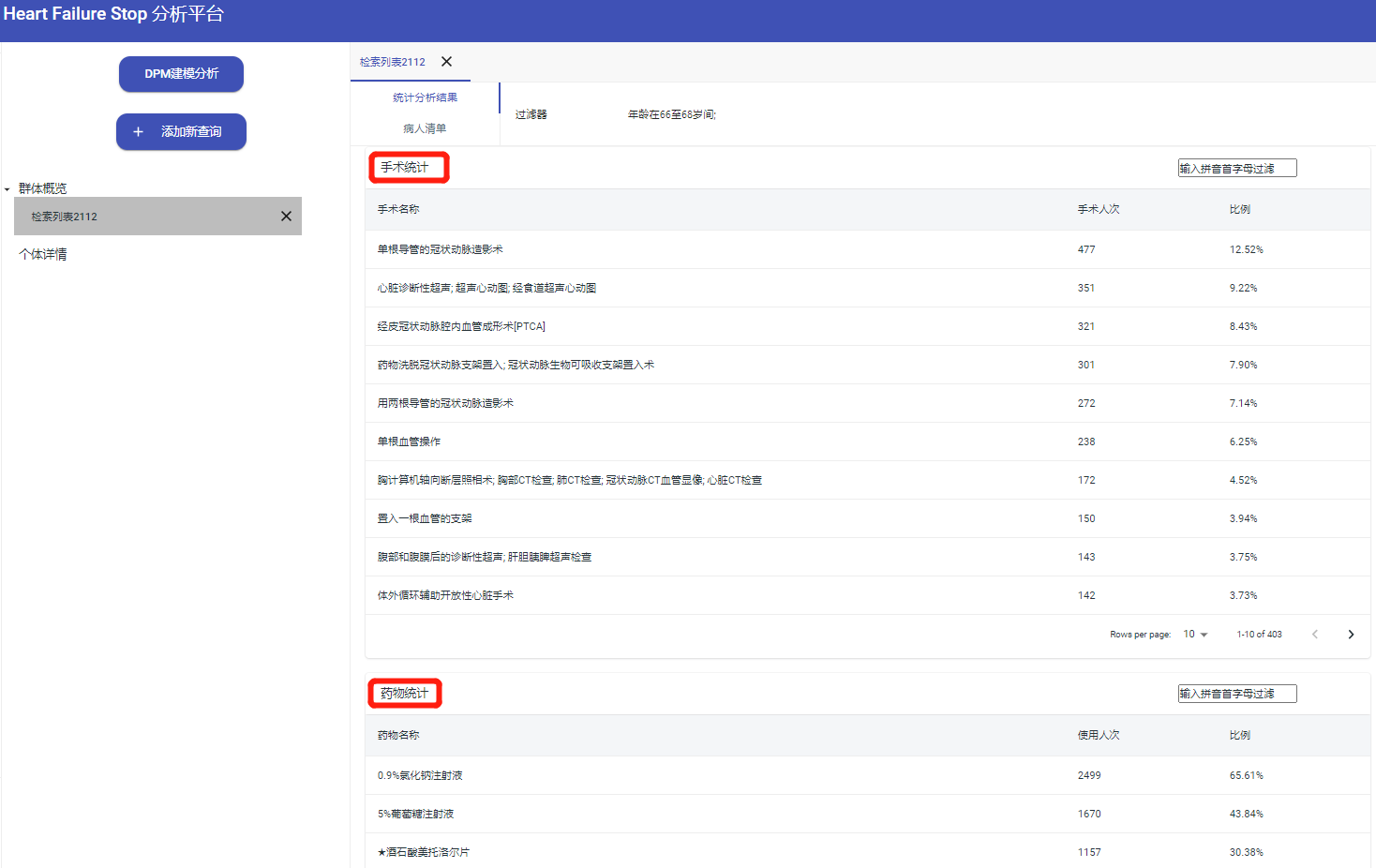


图4.9 患者群体的手术和用药统计

* 个体详情

个体详情是本论文重点想要展示的内容。有两种方式可以进入到个体详情页面：一是在主界面选择查询类型的时候，选择“个体分析”，然后输入病人的ID来查找特定的患者，此时页面上就会显示该患者的详细信息，如图4.10所示；二是在前面的患者群体列表中，选择一个患者，就会跳转到该患者的详情页面，效果与图4.10相同。该页面首先会显示患者个体的基本信息和历次入院对应的状态转移曲线。在下方的病人入院轨迹中，会详细说明每次入院的时间、症状、诊断及手术等信息，默认选中患者的最后一次入院，医护人员可以通过点击入院轨迹上的节点切换到其他入院。图4.11为该患者在对应入院期间的实验室检查和各项体征生理指标，医护人员可以在左侧的列表中选择感兴趣的因素进行查看和分析。图4.12则为患者在该次入院期间的医嘱和各项检查信息，同样可以在左侧进行切换。

## 本章小结

本章对本论文提出的基于贝叶斯隐马尔可夫模型的心衰演变过程分析方法进行了可视化系统实现。首先介绍了系统的整体框架和功能模块，主要包括群体统计和个体分析两大模块。接着展示了各个功能模块的系统界面，本系统页面直观简洁，便于临床医护人员操作和理解。



图4.10 患者个体的基本信息与入院轨迹

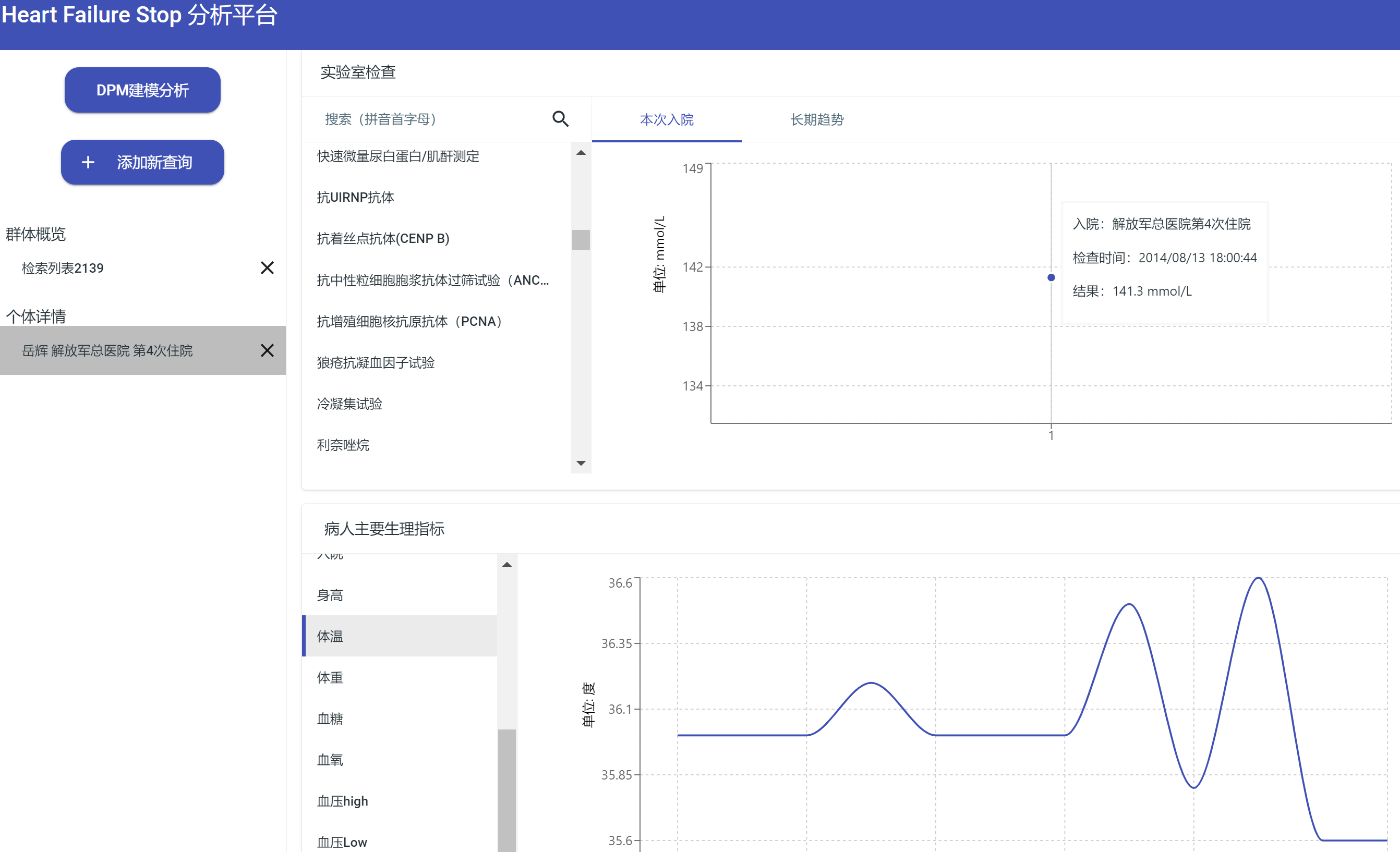


图4.11 患者个体的实验室检查和生理指标

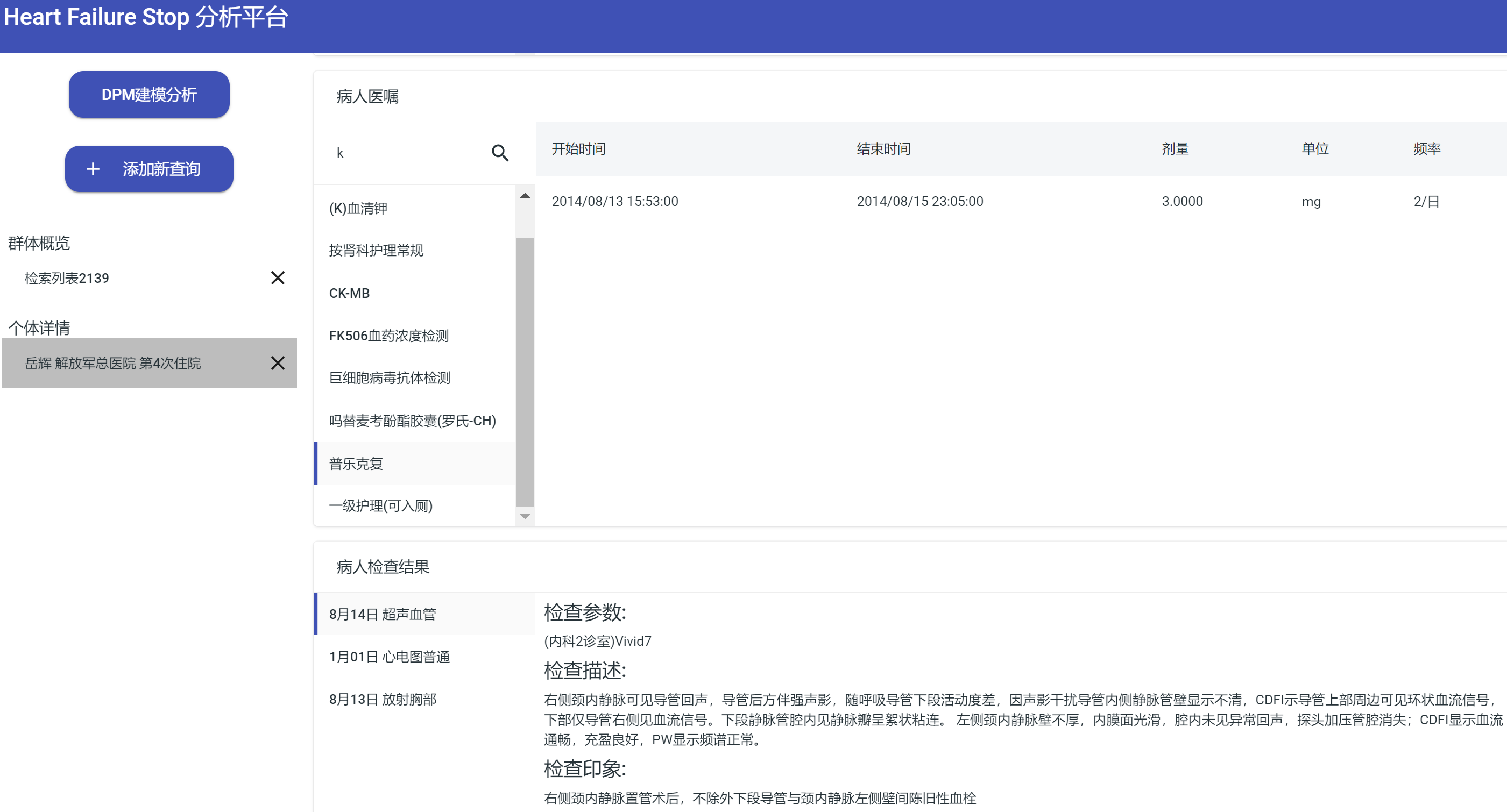


图4.12 患者个体的医嘱和各项检查信息

# 总结与展望

## 总结

心力衰竭，作为世界范围内致死率最高的心血管疾病的最终发展阶段，是一个严重的全球性公众健康问题，且治疗费用昂贵，给患者带来了沉重的经济负担和精神压力。对心衰的演变过程进行分析，挖掘其发展变化的规律，有助于及早发现患者的心衰演变轨迹，预测将来的临床表现，从而帮助医生制定相应的治疗计划，提供更高效的医疗服务和全程管理。

心衰等慢性病的演变过程分析由于其重大的临床意义得到了广泛的关注和研究。然而传统的临床队列研究存在着诸如入组出组标准过于严格使得结果不具有普适性、患者特征项纳入不全以及无法处理纵向数据缺失等问题。随着各大医院信息化建设进程的不断推进，海量的患者诊疗信息保存在了EHR中，为心衰演变过程分析提供了宝贵的数据来源。基于EHR的疾病演变过程建模可以克服队列研究的缺点，然而依然存在着仅考虑一种单一的演变模式以及对EHR数据利用不充分的问题。针对这些问题，本论文提出了一种基于贝叶斯隐马尔可夫模型的概率生成模型，用来对心衰演变过程进行建模。

在正式建模之前，首先对本论文采用的心衰患者EHR数据集进行了统计分析。通过对基本信息、实验室检查信息、诊断信息、用药信息及手术信息等患者数据进行分类统计，得到了患者各项特征的基本分布情况。为了初步探究心衰演变过程的特点，选取了三名患者的历次入院记录进行了详细的分析。

建模的思想是将患者每次入院期间产生的一系列临床观察项看作是隐马尔可夫过程中的观测序列，它们对应着一个隐藏的心衰状态。训练模型的过程就是根据现有的EHR数据集推导出各个心衰状态下具体的临床观察项分布概率，并推导出每次入院对应的心衰状态，据此对患者的心衰演变过程进行分析。最后还提出了一种定量评估方案来评估该模型的有效性和准确性。

实验结果表明，基于贝叶斯隐马尔可夫模型的心衰演变过程建模方法能够发掘出隐藏在EHR数据集中的心衰状态，进而很好地描述和代表心衰患者的病情发展轨迹。相比基本的朴素贝叶斯模型，本论文提出的模型在预测患者将来的临床表现上准确率更高。

在上述工作的基础上，本论文设计并实现了心衰患者数据分析系统。该系统主要包括群体统计和个体分析两大模块，可以帮助医生迅速定位感兴趣的患者群体，并在线对患者个体进行心衰演变过程分析。因此，该系统可以辅助临床医生及时制定治疗计划，提高医疗服务的质量。

## 展望

本论文提出了一种基于贝叶斯隐马尔可夫模型的概率生成模型，通过学习大量的EHR数据来分析患者的心衰演变过程，取得了较好的实验结果，但仍存在一定的局限性，可以通过以下几个方面来改进和扩展：

* 本论文相较于其他研究纳入了更多更综合的患者特征项，但仍然不够全面。下一步的工作可以考虑将基因组学信息、心电图信息等纳入心衰演变过程分析模型，使建模结果更加准确。
* 由于缺乏适当的金标准[38]，很难定量地去评估本论文提出的模型的性能。本论文提出了一种临床表现预测准确率的评估方案，并与基准模型朴素贝叶斯模型进行对比。将来还可以进一步设计评估实验，通过经验丰富的临床医生打分来进行评估。
* 本论文提出的模型目前只是发掘出了几种隐藏的心衰状态，允许患者沿着任何状态转移路线发展。将来可以在此基础上进一步地进行心衰演变轨迹聚类，提取出几种常见的心衰演变模式。
* 将来还可以进一步研究如何从大量的原始病历数据中挖掘出医疗干预影响疾病进展的关联模式和知识，针对个体患者推荐适合的干预手段，以阻碍或延缓疾病的发展，降低临床主要不良事件的发生风险

# 参考文献

1. Organization W H. Global status report on noncommunicable disease. [EB/OL].

https://www.who.int/topics/noncommunicable\_diseases/en/.

1. World Health Statistics 2020: Monitoring health for the SDGs. Retrieved May 18, 2020
2. JD Stanaway, A Afshin, E Gakidou, et al. Global, regional, and national comparative risk assessment of 84 behavioural, environmental and occupational, and metabolic risks or clusters of risks for 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. Lancet, 392 (2018), pp. 1923-1994
3. Virani S S , Alonso A , Benjamin E J , et al. Heart Disease and Stroke Statistics—2020 Update: A Report From the American Heart Association[J]. Circulation, 2020, 141(9):CIR0000000000000757-.
4. Organization W H. Global status report on cardiovascular disease. [EB/OL].

https://www.who.int/cardiovascular\_diseases/about\_cvd/zh/.

1. Palmiero P . Editorial -World Health Organization Cardiovascular Disease Risk Charts: Revised Models to Estimate Risk in 21 Global Regions[J]. World Heart Journal, 2019, 11(4):261.
2. 陆永崇. 心力衰竭诊断与治疗研究进展[J]. 医学理论与实践, 2015, 000(013):641-645.
3. 中华医学会心血管病学分会, 中华心血管病杂志编辑委员会. 慢性心力衰竭诊断治疗指南[J]. 中华心血管病杂志, 2007, 035(012):1076-1095.
4. 戴闺柱. 慢性心力衰竭治疗的现代概念[J]. 中华心血管病杂志, 2000, 28(1).
5. 郭静萱, 李海燕. 慢性心力衰竭的诊治进展[J]. 中国实用内科杂志, 2007(1):11-13.
6. D. Mould. Models for disease progression: new approaches and uses. Clinical Pharmacology & TEHRapeutics, 92(1):125-131, 2012.
7. Huang Z , Dong W , Duan H . A probabilistic topic model for clinical risk stratification from electronic health records[J]. Journal of Biomedical Informatics, 2015, 58.
8. Huang Z , Dong W , Duan H , et al. A regularized deep learning approach for clinical risk prediction of acute coronary syndrome using electronic health records[J]. IEEE transactions on bio-medical engineering, 2017:1-1.
9. Zhengxing H , Zhenxiao G , Wei D , et al. Relational regularized risk prediction of acute coronary syndrome using electronic health records[J]. Information Sciences, 2018, 465:S0020025516320941-.
10. Huang Z , Chan T M , Dong W . MACE prediction of acute coronary syndrome via boosted resampling classification using electronic medical records[J]. Journal of Biomedical Informatics, 2017, 66:161-170.
11. Yin L , Huang Z , Dong W , et al. Utilizing Electronic Medical Records to Discover Changing Trends of Medical Behaviors Over Time[J]. Methods of Information in Medicine, 2017, 56(S 01):e49-e66.
12. P.B. Jensen, L.J. Jensen, S. Brunak. Mining electronic health records: towards better research applications and clinical care. Nat Rev Genet, 13 (6):395–405, 2012.
13. 沈洪兵、齐秀英主编．流行病学（第8版）．北京：人民卫生出版社，2013.3：54.
14. 胡丹青. 基于电子病历的急性冠脉综合征患者主要不良心血管事件预测[D]. 2017.
15. Richard K. Burt, et al. Effect of Nonmyeloablative Hematopoietic Stem Cell Transplantation vs Continued Disease-Modifying TEHRapy on Disease Progression in Patients with Relapsing-Remitting Multiple Sclerosis: A Randomized Clinical Trial. JAMA, 321(2):165-174, 2019
16. L. Chen, et al. A novel genotype-based clinicopathology classification of arrhythmogenic cardiomyopathy provides novel insights into disease progression. European Heart Journal, doi.10.1093/eurheartj/ehz172, 2019
17. Ahnul Ha, et al. Association of Angle Width With Progression of Normal-Tension Glaucoma, JAMA Ophthalmology, 137(1):13-20, 2019
18. Eduardo De Pablo-Fernandez, Raph Goldacre, Julia Pakpoor, Alastair J. Noyce, Thomas T. Warner. Association between diabetes and subsequent Parkinson disease, Neurology, 91(2):e139-e142, 2018.
19. K. Ito, S. Ahadieh, B. Corrigan, J. French, T. Fullerton, and T. Tensfeldt. Disease progression meta-analysis model in alzheimer's disease. Alzheimer's & Dementia, 6(1):39-53, 2010
20. Yan A T , Jong P , Yan R T , et al. Clinical trial--derived risk model may not generalize to real-world patients with acute coronary syndrome.[J]. American Heart Journal, 2004, 148(6):1020-1027.
21. 李玉亮. 血清学指标在心衰患者NYHA心功能分级中的诊断意义[J]. 湖南师范大学学报:医学版, 2015.
22. Paredes S , Rocha T , Carvalho P D , et al. Long term cardiovascular risk models' combination[J]. Computer Methods & Programs in Biomedicine, 2011, 101(3):231-242.
23. Cooney M T , Dudina A L , Graham I M . Value and Limitations of Existing Scores for the Assessment of Cardiovascular Risk. A Review for Clinicians[J]. Journal of the American College of Cardiology, 2009, 54(14):1209-1227.
24. Jonnagaddala J, Liaw S, Ran P, et at. Coronary artery disease risk assessment from unstructured electronic health records using text mining[J]. Journal of Biomedical Informatics, 2015, 58:S203-S210.
25. Z. Huang, W. Dong, P. Bath, L. Ji, and H. Duan. On mining latent treatment patterns from electronic medical records. Data Mining and Knowledge Discovery, 2015, 29(4):914-949
26. Panigrahi S , Kundu A , Sural S , et al. Credit card fraud detection: A fusion approach using Dempster–Shafer theory and Bayesian learning[J]. Information Fusion, 2009, 10(4):354-363.
27. Wang J , Hu Y , Xiao F , et al. A novel method to use fuzzy soft sets in decision making based on ambiguity measure and Dempster-Shafer theory of evidence: An application in medical diagnosis[J]. Artificial Intelligence in Medicine, 2016, 69(may):1-11.
28. Dutta, Palash. Uncertainty Modeling in Risk Assessment Based on Dempster–Shafer Theory of Evidence with Generalized Fuzzy Focal Elements[J]. Fuzzy Information & Engineering, 2015, 7( 1):15-30.
29. M. Cohen, A. Grossman, D. Morabito, M. M. Knudson, A. Butte, and G. Manley. Identification of complex metabolic states in critically injured patients using bioinformatic cluster analysis. Critical Care, 14(1):R10, 2010.
30. J. Zhou, J. Liu, V. A. Narayan, and J. Ye. Modeling disease progression via fused sparse group lasso. In KDD, pages 1095-1103. ACM, 2012.
31. J. Zhou, J. Liu, V. A. Narayan, and J. Ye. Modeling disease progression via multi-task learning. NeuroImage, 78(0):233-248, 2013
32. X. Wang, D. Sontag and F. Wang. Unsupervised learning of disease progression models, KDD’14, August 24-27, 2014, New York, NY, USA
33. Z. Sun, S. Ghosh, Y. Li, Y. Cheng, A. Mohan, C. Sampaio, and J. Hu. A probabilistic disease progression modeling approach and its application to integrated Huntington’s disease observational data, JAIMA Open, 2(1):123-130, 2019
34. C. H. Jackson, L. D. Sharples, S. G. Thompson, S. W. Duffy, and E. Couto. Multistate markov models for disease progression with classification error. Journal of the Royal Statistical Society: Series D (The Statistician), 52(2):193-209, 2003.
35. R. Sukkar, E. Katz, Y. Zhang, D. Raunig, and B. T. Wyman. Disease progression modeling using hidden markov models. In EMBC, pages 2845-2848. IEEE, 2012
36. Goldwater S , Griffiths T . A fully Bayesian approach to unsupervised part-of-speech tagging[C]// ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, June 23-30, 2007, Prague, Czech Republic. DBLP, 2007.
37. Baoxing, Huai, Enhong, et al. Toward Personalized Context Recognition for Mobile Users: A Semisupervised Bayesian HMM Approach[J]. Acm Transactions on Knowledge Discovery from Data, 2014.
38. Liechty J , Pieters R , Wedel M . Global and local covert visual attention: Evidence from a bayesian hidden markov model[J]. Psychometrika, 2003, 68(4):519-541.
39. Guha S , Li Y , Neuberg D . Bayesian Hidden Markov Modeling of Array CGH Data[J]. Journal of the American Statal Association, 2008, 103(jun.):485-497.
40. 杨欣, 任斌辉, 许林. 脑利钠肽BNP的临床应用进展[J]. 现代生物医学进展, 2013(04):775-777.
41. 蔡文娟, 李洋, 门岚,等. 尿微量白蛋白/肌酐比值对糖尿病肾病早期诊断价值[J]. 中国实验诊断学, 2013, 17(009):1650-1653.
42. 李晓阳, 郭学青. C反应蛋白[J]. 河北医学, 2002, 8(9):856-857.
43. 王晶, 刘晶晶, 刘金凤,等. 血浆D-二聚体水平与急性脑梗死患者预后的关系[J]. 临床神经病学杂志, 2014, 27(002):123-124.
44. Nardino R J , Vender R J , Herbert P N . Overuse of acid-suppressive therapy in hospitalized patients[J]. American Journal of Gastroenterology, 2000, 95(11):3118-3122.
45. 佚名. β受体阻滞剂[J]. 医学理论与实践(10):1218-1219.
46. 乔文豪, 王秀, 王佳丽,等. 抗血小板药物药理作用和临床应用的研究进展[J]. 安徽医药, 2014, 000(009):1621-1625.
47. 张茹岩. 他汀类药物在急性冠状动脉综合征的应用[J]. 养生保健指南, 2017, 000(005):1.
48. 李占全, 金元哲. 冠状动脉造影与临床 第3版[M]. 辽宁科学技术出版社, 2012.
49. 颜勇, 杨庭树, 苗冬梅,等. 再血管化治疗对急性心肌梗死患者血浆激素水平和心功能的影响[J]. 中国循环杂志, 2004, 19(001):31-33.
50. 杨娅. 超声掌中宝 心血管系统[M]// 超声掌中宝, 心血管系统. 科学技术文献出版社, 2009.
51. 陈婵, 张鹏, 王娟,等. 射血分数保留和射血分数减低心力衰竭中医四诊及证候特征研究[J]. 中西医结合心脑血管病杂志, 2019, 17(07):16-19.
52. J.Pearl. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference, Morgan Kaufmann PublisEHRs, San Mateo, Calif., 1988
53. Cheng J , Greiner R , Kelly J , et al. Learning Bayesian Networks from Data: An Information-Theory BasedApproach[J]. Artificial Intelligence, 2002, 137(1-2):43-90.
54. 张连文郭海鹏. 贝叶斯网引论[M]. 科学出版社, 2006.
55. Terent'Yev A N , Bidyuk P I . Method of probabilistic inference from learning data in Bayesian networks[J]. Cybernetics & Systems Analysis, 2007, 43(3):391-396.
56. José A. Gámez, Antonio Salmerón. Probabilistic graphical models[J]. International Journal of Intelligent Systems, 2003, 18(2):149-151.
57. Mendes E , Mosley N . Bayesian Network Models for Web Effort Prediction: A Comparative Study[J]. IEEE Transactions on Software Engineering, 2008, 34(6):723-737.
58. Kapitonova Y V , Mishchenko N M , Felizhanko O D , et al. USING BAYESIAN NETWORKS FOR MONITORING COMPUTER USERS[J]. Cybernetics & Systems Analysis, 2004, 40(6):p.789-799.
59. Ben-Gal I . Bayesian Networks[M]// Encyclopedia of Statistics in Quality and Reliability. John Wiley & Sons, Ltd, 2008.
60. Niculescu R S , Mitchell T M , Rao R B . Bayesian Network Learning with Parameter Constraints[J]. Journal of Machine Learning Research, 2006, 7(3):1357-1383.
61. JamesO.Berger, 贾乃光, 吴喜之,等. 《统计决策论及贝叶斯分析》[J]. 数理统计与管理, 2000, 19(2):50-50.
62. 陈希孺. 数理统计学简史[M]. 湖南教育出版社, 2002.
63. Bidyuk P I , A.N. Terent’ev. Construction and Methods of Learning of Bayesian Networks[M]. Kluwer Academic Publishers, 2005.
64. Baum L E , Petrie T . Statistical Inference for Probabilistic Functions of Finite State Markov Chains[J]. Annals of Mathematical Stats, 1966, 37(6):1554-1563.
65. Baum L E , Eagon J A . An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology[J]. Bull.amer.math.stat, 1967, 37(3):360-363.
66. Baum L E , Petrie T , Soules G , et al. A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains[J]. Annals of Mathematical Stats, 1970, 41(1):164-171.
67. 朱明, 郭春生. 隐马尔可夫模型及其最新应用与发展[J]. 计算机系统应用, 2010, 19(7):255-259.
68. 黄岗. 马尔可夫及隐马尔可夫模型的应用[J]. 电子设计工程, 2013, 21(017):60-62.
69. Norris J R . Markov Chains[M]. DBLP, 1998.
70. Odentrantz, John. Markov Chains: Gibbs Fields, Monte Carlo Simulation, and Queues[J]. Technometrics, 2010, 42(4):438-439.
71. byRichardSerfozo. Basics of Applied Stochastic Processes[M]. Springer Berlin Heidelberg, 2009.
72. Jacobsen M . Markov Chains and Markov Processes[M]// Encyclopedia of Actuarial Science. American Cancer Society, 2015.
73. Rabiner L R . A tutorial on hidden Markov models and selected applications in speech recognition[J]. Proc IEEE, 1989, 77.
74. 莫玉婷. 基于贝叶斯分析的隐马尔可夫模型及其应用[D].华南农业大学,2016.
75. 万稚慧, 刘金山. 基于隐马尔可夫模型的array-CGH数据贝叶斯分析[J]. 系统工程理论与实践, 2016(3):684-693.
76. 裘亚峥, 任叶庆, 刘诚. 概率论与数理统计.第4版[M]. 科学出版社, 2015.
77. George, Casella, &, et al. Explaining the Gibbs Sampler[J]. American Statistician, 1992.
78. Robert C P . Estimation of finite mixture distributions through Bayesian sampling[J]. Journal of the Royal Statal Society, 1994, 56(2):363-375.
79. 安西, 祐一郎. Pattern Recognition and Machine Learning[M]// Stat Sci. Springer, 2006.
80. Gelfand A E , Smith A F M . Sampling-Based Approaches to Calculating Marginal Densities[J]. Publications of the American Statal Association, 1990, 85:398-409.
81. 徐戈,王厚峰.自然语言处理中主题模型的发展[J].计算机学报,2011,34(08):1423-1436.
82. Blei D M , Ng A Y , Jordan M I , et al. Latent Dirichlet Allocation[J]. J. Mach. Learn. Res, 2012, 3:993-1022.
83. Griffiths T L , Steyvers M . Finding scientific topics[J]. Proceedings of the National Academy of Sciences of the United States of America, 2004, 101(Supplement 1):p.5228-5235.
84. Teh Y W , Jordan M I , Beal M J , et al. Hierarchical Dirichlet Processes[J]. Journal of the American Statal Association, 2006, 101(December):1566-1581.
85. Russell, Stuart J . Artificial Intelligence: A Modern Approach[M]. 人民邮电出版社, 2002.
86. Rish I . An empirical study of the naive Bayes classifier[J]. Journal of Universal Computer Science, 2001, 1(2):127.
87. Frank E , Bouckaert R R . Naive Bayes for Text Classification with Unbalanced Classes[C]// European Conference on Principles of Data Mining and Knowledge Discovery. Springer, Berlin, Heidelberg, 2006.

# 作者简介

王慧兰，生于1995年7月，湖北黄冈人，2014年9月进入华中科技大学生命科学与技术学院就读生物医学工程专业，2018年6月获得华中科技大学工学学士学位。同年9月推荐免试到浙江大学生物医学工程与仪器科学学院生物医学信息学方向攻读硕士学位。