分类号：　　密　　级：公开
学　号：20232114051　　单位代码：10759

# 石河子大学
## 硕士学位论文

# 2005-2024年新疆生产建设兵团某市乙肝流行特征和预测模型研究

学位申请人　李晓雪
指导教师　郭恒　副教授
孙涛　主管医师

申请学位类别　专业硕士
专业名称　公共卫生
研究领域　流行病与卫生统计学
所在学院　公共卫生学院

中国·新疆·石河子
2026年5月

分类号：　　密　　级：公开

学　号：20232114051　　单位代码：10759

# 石河子大学

## 硕士学位论文

# 2005-2024年新疆生产建设兵团某市乙肝流行特征和预测模型研究

| | |
|---|---|
| 学位申请人 | 李晓雪 |
| 指导教师 | 郭恒　副教授 |
| | 孙涛　主管医师 |
| 申请学位类别 | 专业硕士 |
| 专业名称 | 公共卫生 |
| 研究领域 | 流行病与卫生统计学 |
| 所在学院 | 公共卫生学院 |

中国·新疆·石河子

2026年5月

# Research on Epidemiological Characteristics and Prediction Models of Hepatitis B in a City of the Xinjiang Production and Construction Corps from 2005 to 2024

A Dissertation Submitted to

**Shihezi University**

In Partial Fulfillment of the Requirements

for the Degree of

**Master of Public Health**

**By**

**Li Xiaoxue**

**(Epidemiology and Health Statistics)**

Dissertation Supervisor:A/Prof. Guo Heng

Dr. Sun Tao

May,2026

# 石河子大学学位论文独创性声明及使用授权声明

## 学位论文独创性声明

本人所呈交的学位论文是在我导师的指导下进行的研究工作及取得的研究成果。据我所知，除文中已经注明引用的内容外，本论文不包含其他个人已经发表或撰写过的研究成果。对本文的研究做出重要贡献的个人和集体，均已在文中作了明确的说明并表示谢意。

研究生签名：[illegible]　　时间：2026年 5月 20日

## 使用授权声明

本人完全了解石河子大学有关保留、使用学位论文的规定，学校有权保留学位论文并向国家主管部门或指定机构送交论文的电子版和纸质版。有权将学位论文在学校图书馆保存并允许被查阅。有权自行或许可他人将学位论文编入有关数据库提供检索服务。有权将学位论文的标题和摘要汇编出版。保密的学位论文在解密后适用本规定。

研究生签名：[illegible]　　时间：2026 年 5 月 20 日

导师签名：[illegible]　　时间：2026 年5 月20 日

# 摘要

目的：本研究旨在分析 2005-2024 年新疆生产建设兵团某市乙肝的流行特征，识别乙肝发病的高危人群与高发区域；构建自回归移动平均模型（ARIMA）、长短期记忆网络模型（LSTM）、支持向量回归模型（SVR）以及 ARIMA-LSTM、ARIMA-SVR 组合模型，并比较模型的预测效能，筛选出适用于该地区的最佳模型，为该地区乙肝疫情的监测预警与精准防控提供数据支持和决策依据。

方法：采用描述性分析和 Joinpoint 回归模型，对 2005-2024 年该地区乙肝发病的三间分布特征及长期趋势进行分析，并结合 2024 年血清学横断面调查数据描述该地区乙肝感染现状。利用 2005-2024 年月发病率数据，分别构建 ARIMA、LSTM、SVR 三类单一模型，以及 ARIMA-LSTM 和 ARIMA-SVR 的串联与并联组合模型，采用均方误差（MSE）、均方根误差（RMSE）、平均绝对误差（MAE）和平均绝对百分比误差（MAPE）4 个指标综合评估并比较各模型预测效果，筛选适用于该地区的最佳预测模型。

结果：

1. 流行特征：2005-2024 年该地区累计报告乙肝 20330 例，年均发病率为 154.66/10 万。Joinpoint 回归分析显示，发病率呈现先下降后上升的变化趋势，但总体仍以下降为主（AAPC=-7.89%，95%CI: -11.39%～-4.25%）。性别分布方面，男性 13320 例（65.52%），女性 7010 例（34.48%），男女性别比为 1.90:1。职业分布以工人（5122 例，25.20%）、离退人员（4534 例，22.30%）和农民（2908 例，14.30%）为主，三者合计占 61.80%。年龄分布显示，病例主要集中在 30-59 岁人群（12942 例，63.70%），0-9 岁年龄组发病数最少（105 例，0.50%）。地区分布方面，市区报告病例数最多，共 10686 例，占 52.60%。

2. 血清学调查：调查 4782 人，HBsAg 阳性率为 5.00%，HBsAb 阳性率为 56.55%，HBcAb 阳性率为 7.49%，HBeAg 阳性率为 0.23%，HBeAb 阳性率为 3.32%。共检出 18 种乙肝血清学标志物组合模式，其中以 HBsAb 单阳性模式（54.41%）和五项均阴性模式（37.75%）最为常见，合计占 92.16%。HBsAg 阳性者共 239 例，主要以“小三阳”（HBsAg、HBeAb、HBcAb 阳性）以及 HBsAg 与 HBcAb 阳性模式为主。

3. 单一预测模型：LSTM 预测效果最佳（RMSE=3.01，MAE=2.54，MSE=9.07，MAPE=17.32%）；ARIMA 次之（RMSE=5.15，MAE=4.22，MSE=26.50，MAPE=32.66%）；SVR 误差较大（RMSE=5.91，MAE=4.61，MSE=34.94，MAPE=29.46%）。

4. ARIMA-LSTM 系列组合模型：ARIMA-LSTM 系列较单一 ARIMA 模型预测效果得到提升。ARIMA-LSTM 均方根误差倒数并联模型表现最佳（RMSE=2.99，MAE=2.43，MSE=8.96，MAPE=15.06%）；其次为 ARIMA-LSTM 等权重并联模型（RMSE=3.56，MAE=2.79，MSE=12.68，

MAPE=16.98%）；ARIMA-LSTM 串联模型误差较大（RMSE=4.92，MAE=4.35，MSE=24.24，MAPE=27.99%）。

5. ARIMA-SVR 系列组合模型：ARIMA-SVR 系列组合模型表现参差不齐。其中，ARIMA-SVR 串联模型较单一 ARIMA 模型有明显提升（RMSE=3.40，MAE=2.85，MSE=11.54，MAPE=21.91%）；ARIMA-SVR 均方根误差倒数并联模型（RMSE=4.87，MAE=4.03，MSE=23.67，MAPE=27.48%）与 ARIMA-SVR 等权重并联模型（RMSE=5.13，MAE=4.19，MSE=26.29，MAPE=28.00%）误差水平与单一 ARIMA 模型相近，预测效果提升有限。

6. 预测结果：ARIMA-LSTM 均方根误差倒数并联模型预测显示，2025 年、2026 年、2027 年各月发病率范围分别为 7.084/10 万～10.185/10 万、7.235/10 万～8.493/10 万和 8.551/10 万～9.954/10 万。2025-2026 年月发病率预测值整体低于 2024 年水平，2027 年较 2026 年略有回升。

结论：

1. 2005-2024 年新疆生产建设兵团某市乙肝报告发病率总体呈下降趋势，但近年来存在一定回升，应持续加强防控。

2. 男性、30-59 岁、工人、离退人员及市区居民发病人数相对较多，应作为该地区乙肝防控的重点关注人群；同时，该地区人群中仍存在一定比例的乙肝易感者，应进一步加强疫苗接种和重点人群筛查。

3. 在预测模型比较方面，单一模型中 LSTM 预测效果最优，ARIMA 次之，SVR 效果最差；组合模型中，ARIMA-LSTM 系列整体优于 ARIMA-SVR 系列。综合比较所有模型，ARIMA-LSTM 均方根误差倒数并联模型预测效果最佳，在捕捉乙肝发病的非线性特征方面具有明显优势，可为新疆生产建设兵团某市乙肝疫情的动态监测与防控决策提供参考依据。

4. ARIMA-LSTM 均方根误差倒数并联模型预测显示，2025-2026 年该地区乙肝月发病率总体较 2024 年有所下降，但 2027 年发病水平较 2026 年略有回升，提示防控压力尚未完全消除，应持续加强乙肝防控措施。

**关键词：**乙肝；流行特征；ARIMA 模型；长短期记忆网络；支持向量回归

# Abstract

Objective: This study aims to analyze the epidemiological characteristics of hepatitis B in a city of the Xinjiang Production and Construction Corps from 2005 to 2024, identify high-risk populations and high-incidence areas for hepatitis B; construct autoregressive integrated moving average (ARIMA) models, long short-term memory network (LSTM) models, support vector regression (SVR) models, as well as combined ARIMA-LSTM and ARIMA-SVR models, and compare the predictive performance of these models to select the optimal model suitable for the region, providing data support and decision-making basis for the monitoring, early warning, and precise prevention and control of hepatitis B epidemic in the region.

Methods: Descriptive analysis and Joinpoint regression model were used to analyze the three distribution characteristics and long-term trend of hepatitis B incidence in this area from 2005 to 2024, and the serological cross-sectional survey data in 2024 were used to describe the current situation of hepatitis B infection in this area. Using the monthly incidence rate data from 2005 to 2024, three types of single models, ARIMA, LSTM, and SVR, as well as the series and parallel combination models of ARIMA LSTM and ARIMA SVR, were constructed. The mean square error (MSE), root mean square error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE) were used to comprehensively evaluate and compare the prediction effects of each model, and select the best prediction model for the region.

Results:

1. Epidemiological characteristics: From 2005 to 2024, 20330 cases of hepatitis B were reported in this area, with an average annual incidence rate of 154.66/100000. Joinpoint regression analysis showed that the incidence rate showed a trend of decreasing first and then increasing, but it still decreased mainly in general (AAPC=-7.89%, 95% CI: -11.39%~-4.25%). In terms of gender distribution, there were 13320 males (65.52%) and 7010 females (34.48%), with a male to female ratio of 1.90:1. The occupational distribution is mainly composed of workers (5122 cases, 25.20%), retirees (4534 cases, 22.30%), and farmers (2908 cases, 14.30%), accounting for 61.80% in total. The age distribution shows that the cases are mainly concentrated in the 30-59 age group (12942 cases, 63.70%), with the lowest incidence in the 0-9 age group (105 cases, 0.50%). In terms of regional distribution, the urban area reported the highest number of cases, with a total of 10686 cases, accounting for 52.60%.

2. Serological survey: A total of 4782 individuals were surveyed, with HBsAg positivity rate of 5.00%, HBsAb positivity rate of 56.55%, HBcAb positivity rate of 7.49%, HBeAg positivity rate of 0.23%, and HBeAb positivity rate of 3.32%. A total of 18 combination patterns of hepatitis B serological markers were detected, of which the single positive pattern (54.41%) and the five negative patterns (37.75%) of HBsAb

were the most common, accounting for 92.16% in total. There were a total of 239 HBsAg positive cases, mainly characterized by "small three positives" (HBsAg, HBeAb, HBcAb positive) and HBsAg and HBcAb positive patterns.

3. Single prediction model: LSTM achieved the best prediction performance (RMSE=3.01, MAE=2.54, MSE=9.07, MAPE=17.32%); followed by ARIMA (RMSE=5.15, MAE=4.22, MSE=26.50, MAPE=32.66%); SVR had a larger error (RMSE=5.91, MAE=4.61, MSE=34.94, MAPE=29.46%).

4. ARIMA-LSTM series combined models: The prediction performance of the ARIMA-LSTM series has been improved compared to the single ARIMA model. The ARIMA-LSTM inverse root mean square error parallel model performs the best (RMSE=2.99, MAE=2.43, MSE=8.96, MAPE=15.06%); followed by the ARIMA-LSTM equal weight parallel model (RMSE=3.56, MAE=2.79, MSE=12.68, MAPE=16.98%); the ARIMA-LSTM series model has a larger error (RMSE=4.92, MAE=4.35, MSE=24.24, MAPE=27.99%).

5. ARIMA-SVR series The performance of the ARIMA-SVR series combined models is uneven. Among them, the ARIMA-SVR series model shows significant improvement compared to the single ARIMA model (RMSE=3.40, MAE=2.85, MSE=11.54, MAPE=21.91%); the ARIMA-SVR parallel model with the reciprocal of root mean square error (RMSE=4.87, MAE=4.03, MSE=23.67, MAPE=27.48%) and the ARIMA-SVR equal weight parallel model (RMSE=5.13, MAE=4.19, MSE=26.29, MAPE=28.00%) have error levels similar to the single ARIMA model, indicating limited improvement in prediction performance.

6. Prediction results: The parallel model of ARIMA-LSTM root mean square error reciprocal predicts that the incidence rate ranges for each month in 2025, 2026, and 2027 will be 7.084/100,000 to 10.185/100,000, 7.235/100,000 to 8.493/100,000, and 8.551/100,000 to 9.954/100,000, respectively. The overall incidence rate for the three years is expected to decrease compared to 2024, but there will be a slight upward trend in 2027.

Conclusion:

1. The reported incidence rate of hepatitis B in this region has generally shown a downward trend from 2005 to 2024, but there has been a certain resurgence in recent years. Therefore, prevention and control measures should be continuously strengthened.

2. Males aged 30-59, workers, retired personnel, and urban residents have a relatively high incidence of hepatitis B, and should be the key focus groups for hepatitis B prevention and control in this region; meanwhile, there is still a certain proportion of hepatitis B susceptible individuals in the population of this region, and vaccination and screening of key groups should be further strengthened.

3. In terms of prediction model comparison, among single models, LSTM exhibits the best prediction performance, followed by ARIMA, and SVR performs the worst; among combined models, the ARIMA-LSTM series outperforms the ARIMA-SVR series overall. Upon comprehensive comparison of all models, the parallel model with the reciprocal of the root mean square error of ARIMA-LSTM

demonstrates the best prediction performance, exhibiting significant advantages in capturing the nonlinear characteristics of hepatitis B incidence. This provides a reference basis for dynamic monitoring and prevention and control decision-making of hepatitis B epidemic in a city of the Xinjiang Production and Construction Corps.

4. The prediction of the ARIMA-LSTM parallel model with reciprocal root mean square error shows that the monthly incidence rate of hepatitis B in this region from 2025 to 2026 will generally decrease compared to 2024, but the incidence level in 2027 will slightly increase compared to 2026, indicating that the prevention and control pressure has not been completely eliminated, and measures for preventing and controlling hepatitis B should be continuously strengthened.

**Key words:** Hepatitis B; Epidemic characteristics; ARIMA model; Long Short-Term Memory network; Support Vector Regression

# 目录

# 中英文缩略词

| 缩略词 | 英文全称 | 中文全称 |
| --- | --- | --- |
| ACF | Auto-correlation Function | 自相关函数 |
| AIC | Akaike Information Criteria | 赤池信息量准则 |
| APC | Annual Percentage Change | 年度百分比变化 |
| AAPC | Average Annual Percentage Change | 平均年度百分比变化 |
| AR | Autoregressive | 自回归 |
| ARIMA | Autoregressive Integrated Moving Average model | 自回归积分滑动平均模型 |
| BIC | Bayesian Information Criteria | 贝叶斯信息准则 |
| BP | Back Propagation Neural Network | BP 神经网络 |
| HBV | Hepatitis B Virus | 乙型肝炎病毒 |
| HBeAb | Hepatitis B e Antibody | 乙型肝炎 e 抗体 |
| HBeAg | Hepatitis B e Antigen | 乙型肝炎 e 抗原 |
| HBsAb | Hepatitis B surface Antibody | 乙型肝炎表面抗体 |
| HBsAg | Hepatitis B surface Antigen | 乙型肝炎表面抗原 |
| HBcAb | Hepatitis B core Antibody | 乙型肝炎核心抗体 |
| IARC | International Agency for Research on Cancer | 国际癌症研究机构 |
| LSTM | Long Short-Term Memory | 长短期记忆网络 |
| MA | Moving Average | 移动平均 |
| MAE | Mean Absolute Error | 平均绝对误差 |
| MAPE | Mean Absolute Percentage Error | 平均绝对百分比误差 |
| ML | Machine Learning | 机器学习 |
| MSE | Mean Square Error | 均方误差 |
| PACF | Partial Auto-correlation Function | 偏自相关函数 |
| RMSE | Root Mean Square Error | 均方根误差 |
| RNN | Recurrent Neural Network | 循环神经网络 |
| SVM | Support Vector Machine | 支持向量机 |
| SVR | Support Vector Regression | 支持向量回归机 |
| WHO | World Health Organization | 世界卫生组织 |
| 95%CI | 95% Confidence Interval | 95%置信区间 |

# 第 1 章 引言

## 1.1 乙肝的流行现状与疾病负担

乙型病毒性肝炎（Hepatitis B，HB），是由乙型肝炎病毒（Hepatitis B Virus，HBV）感染形成的一种以肝脏炎症病变为主，并可以引起全身多器官损伤的传染病，主要通过血液传播、性传播和母婴垂直传播来威胁人类的健康，其中乙肝患者和 HBV 携带者是乙肝的主要传染源[1-3]。目前乙肝已经演变成全球性的重大公共卫生问题。根据世界卫生组织（World Health Organization，WHO）估计，截至 2022 年，全球范围内有 2.54 亿人慢性乙肝感染，每年有 120 万新发感染者，2022 年因乙肝死亡人数达到 110 万人，主要死因为肝硬化和肝细胞癌，其中西太平洋区域和非洲区域的感染者最多[4]。中国作为西太平洋区域人口最多的国家之一，亦是乙肝的高发国家[5]。根据国际癌症研究机构（International Agency for Research on Cancer，IARC）评估认定，HBV 是七种Ⅰ类人类致癌物的病毒之一，慢性感染后容易导致肝硬化，进而发展为肝细胞癌[6, 7]。近年来，肝细胞癌的患病率呈上升趋势，全球范围内由 HBV 感染导致的肝细胞癌占 45%[8]。根据中国国家癌症登记数据，约 83.2%的肝细胞癌死亡病例归因于已知危险因素，其中 77.7%-88.0%归因于 HBV 和/或丙型肝炎病毒感染[9]。

自我国从 1992 年开始将乙肝疫苗纳入计划免疫管理，正式实施乙肝疫苗接种，我国乙肝的传播与流行得到了一定的控制，已经从高度流行降至目前的中度流行[10, 11]。但由于我国的乙肝患者人口基数较大，而乙肝患病较为隐匿、病程较长、难以治愈，我国现存的 HBV 感染者数量众多，HBV 感染情况依然较为严重[12]。近期研究表明，中国是全球乙肝负担最重的国家，中国约有 7500 万人患有乙肝，感染病例占全球慢性乙肝病例的三分之一[13]。肝细胞癌作为乙肝极为严重的并发症之一，2022 年全球约有 86.5 万例肝癌新发病例，75.8 万例死亡病例，其中发病和死亡负担主要集中在东亚地区，中国约占全球病例的近一半[14]。我国乙肝感染负担较重，是全球消除 HBV 感染进程中面临的重要挑战之一，预计到 2030 年，我国的乙肝流行状况将在全球消除乙肝目标实现过程中发挥关键影响[3, 15]。此外，流行病学证据表明，中国乙型肝炎发病存在明显的区域差异和空间聚集现象，高发区域主要集中在西北地区。其中新疆、青海和甘肃等省份属于乙肝高发区域，平均发病率明显高于东部地区[16]。最新研究显示，2007-2023 年我国乙肝发病率总体呈下降趋势，但区域差异依然明显，其中新疆地区发病率由 2007 年的 118.3/10 万下降至 2023 年的 89.4/10 万，仍高于全国平均水平（64.3/10 万）[17]。

## 1.2 发病预测模型的研究现状

对于乙肝防控而言，精准预测其发病趋势至关重要。有效的预测能为乙肝的防控提供数据支撑，及时预防和控制乙肝的流行，节省控制乙肝流行的时间和费用，为乙肝防控提供个性化的防控措施，以达到预防和控制乙肝的目的[18]。乙肝的预测模型主要通过建立数学模型来预测流行趋势，为防控策略制定提供参考[19]。目前常见传染病预测模型的类型主要包括传统的时间序列分析、机器学习算法以及不同组合模型等。

传统的时间序列分析模型是应用数学模型预测未来发展趋势的一种方法。自回归移动平均模型（Autoregressive Integrated Moving Average model，ARIMA 模型）是一种应用较多的经典的时间序列分析模型，由 Box 和 Jenkins 联合提出，该模型仅将历史观测值作为分析对象，不考虑其他外部因素的作用，以时间来替代各种影响因素的综合效应[20]。ARIMA 模型的参数设定较为简便，并且能够确保较高的预测准确性，是时间序列预测中最为有效的线性模型之一，广泛应用于传染病预测[21, 22]。

近年来，机器学习（Machine Learning，ML）算法在传染病发病预测领域受到越来越多的关注。目前，乙肝预测中常用的机器学习模型包括 BP 神经网络（Back Propagation Neural Network，BP）、长短期记忆网络（Long Short-Term Memory，LSTM）、循环神经网络（Recurrent Neural Network，RNN）和支持向量机（Support Vector Machine，SVM）等。根据模型结构和学习方式的不同，机器学习方法通常可以分为传统机器学习模型（如 SVM、BP 神经网络）和深度学习模型（如 RNN、LSTM）两大类[23]。LSTM 是 RNN 的一种特殊形式，它通过引入门控机制有效地解决了 RNN 中的梯度消失和长期依赖学习问题，使其在分析时间序列数据方面表现更为出色[24]。目前，已经有学者在传染病预测方面验证了 LSTM 模型的有效性，该模型能够较好地进行时间序列的非线性趋势预测，且有较高的预测效能[25, 26]。SVM 是利用核技巧将低维空间线性不可分的样本映射到高维空间，实现样本线性可分的一种机器学习模型。其凭借在小样本、非线性以及高维模式识别问题中的较好表现，受到愈来愈多研究者的重视[27]。SVM 模型按用途可分为支持向量分类机（Support Vector Classification，SVC）和支持向量回归机（Support Vector Regression，SVR）两类，其中 SVR 模型应用于连续性数据的回归预测，在处理小样本、非线性及高维数据的时间序列预测任务中具有一定优势[28, 29]。

智能组合模型的基本原理在于将两个或两个以上的单一预测模型，运用特定方法加以结合，从而构建出一个更具稳定性与可信度的新模型，以此提升模型的预测效果[30, 31]。多数研究表明，合理构建的组合模型能够在一定程度上整合单一模型的优势，从而有望降低预测误差并提高预测效果[32, 33]。目前，较为常见的组合模型包括 ARIMA-LSTM 组合模型和 ARIMA-SVR 组合模型等。ARIMA 模型在捕捉线性趋势方面具有较好的稳定性，而 LSTM 和 SVR 分别在非线性关系挖掘和小样本高维数据建模方面各具优势。通

过将传统时间序列模型与不同类型的机器学习模型进行组合，可以在一定程度上实现线性建模与非线性建模能力的优势互补，为传染病监测与预警提供更加可靠的技术支持。

不同的传染病类型适合不同的预测模型，国内外已有大量的预测模型用来预测多种传染病的发病趋势。Zhao 等[34]在预测全球新冠肺炎新发病例时发现，ARIMA 模型的预测效果优于 MLR 和 Prophet 模型。杨敏雪等[35]利用乌鲁木齐市 2012 年至 2021 年的乙肝发病数据建立 LSTM 模型和 ARIMA 预测模型，结果发现，LSTM 的模型预测效果优于 ARIMA。GUO 等[36]分别构建了 SVM、ARIMA 和 LSTM，结果发现非线性模型（SVM、LSTM）优于线性模型（ARIMA），LSTM 的预测精度更高，最适合预测戊型肝炎的月发病率和病例数。Zhang 等[37]比较了 ARIMA 与 LSTM 模型在不同时间尺度下对中国出血热发病率的预测效果，发现不同模型适合不同时间尺度的预测，其中 ARIMA 在月度预测上的误差低于 LSTM。

近年来，组合模型在传染病预测领域的应用日益广泛，在手足口病、艾滋病、副伤寒、肺结核以及流感等多种传染病的预测中均有涉及[38-40]。刘今等[41]应用 ARIMA-BP 神经网络组合模型和 ARIMA-LSTM 组合模型对河南省手足口病的发病情况进行预测，经过比较发现 ARIMA-LSTM 组合模型对手足口病的预测效果更好。Chen 等[42]构建了 ARIMA 和 LSTM 及其组合模型用于预测东亚地区艾滋病发病和死亡趋势，其组合模型在预测效果上优于单一 ARIMA 和 LSTM 模型。而在一项针对埃塞俄比亚结核病发病率的预测研究中，对比 ARIMA、LSTM 及 ARIMA-LSTM 三种模型的预测效果，结果显示单一 LSTM 模型反而表现更佳[43]。另一研究运用 SARIMA 模型、LSTM 模型以及二者组合模型对西藏某医院肺结核病病例进行预测，也发现由于干扰因素以及组合模型的方法不同等原因，单一 LSTM 模型预测效果优于 SARIM 模型和组合模型[44]。然而，组合模型对于乙肝的发病预测研究较少，其预测价值亟待探讨。

## 1.3 新疆生产建设兵团乙肝流行与预测模型现状

新疆生产建设兵团地处西北内陆，辖区内气候与地理条件差异较大，各民族生活方式与行为习惯亦不尽相同，导致该地区乙肝的流行特征呈现出多样性。新疆生产建设兵团作为我国乙肝的高发地区，发病率常年高于我国平均发病水平。2010-2021 年，新疆生产建设兵团乙肝报告病例数达 40675 例，年均发病率为 119.4/10 万[45]，而同期中国共报告乙型肝炎病例 900 多万例（2010-2018 年），年均发病率 75.93/10 万[46]，说明兵团地区乙肝发病负担显著高于全国平均水平。

新疆生产建设兵团目前缺乏针对性、系统性和长期性的乙肝流行病学研究，关于乙肝发病率预测模型的相关研究也较为欠缺，且目前乙肝发病预测模型大多集中于经典的时间序列分析（如 ARIMA 模型），而基于机器学习算法的预测模型研究相对不足，尤

其深度学习（如 LSTM）与传统机器学习（如 SVR）方法的对比应用较为匮乏。此外，ARIMA 与上述机器学习算法的组合模型研究亦较少。ARIMA 模型、LSTM 模型、SVR 模型以及 ARIMA 模型与两者的组合模型对于乙肝的预测效果比较研究亟需进一步探讨。

新疆生产建设兵团某市作为新疆生产建设兵团人口规模最大的城市，辖区包括市区和十四个团场，人口总数为 76.11 万人，约占新疆生产建设兵团总人口的四分之一，其人口规模、生活方式以及居住环境能够较为准确地反映新疆生产建设兵团的整体情况。因此，对该地区乙肝流行特征进行分析，并对其发病预测模型进行研究，对于新疆生产建设兵团制定针对性的预防控制措施、评价防控效果、及时发出预警信号，以及保障居民健康具有重要意义。

## 1.4 研究目的与意义

本研究基于新疆生产建设兵团某市 2005-2024 年乙肝发病数据，分析该地区乙肝的流行特征，深入了解该地区乙肝的流行规律和发展趋势，识别高发区域和易感人群；通过 Joinpoint 回归模型分析乙肝发病率的时间变化特点；并结合 2024 年血清学横断面调查数据描述该地区乙肝感染现状。同时，分别构建 ARIMA 模型、LSTM 模型、ARIMA-LSTM 组合模型、SVR 模型和 ARIMA-SVR 组合模型，评估模型的预测效果，挑选出最适宜该地区的乙肝发病率预测模型并进行预测。通过对乙肝流行特征和预测模型的研究，本研究旨在为新疆生产建设兵团预防和控制乙肝流行提供理论基础和数据支撑，明确防控重点，并为制定有针对性、实用性的干预方法和防控策略提供可靠依据。