·流行病与统计方法·

基于百度指数对青海省流感预测的机器学习模型比较

汪晖[1],沈蕊[3],张明慧[1],靳海晓[2],王朝才[2],冯录召[3],徐莉立[2]
1.青海大学医学院公共卫生系,青海 西宁 810001;
2.青海省疾病预防控制中心传染病预防控制所,青海 西宁 810000;3.北京协和医学院群医学及公共卫生学院

摘要:目的　分析青海省流感与百度指数之间的相关性及时滞变化特征,并建立适用于青海省的流感预测模型。**方法**　“中国流感监测信息系统”获得 2015—2024 年青海省流感病毒阳性率监测数据,并挖掘与流感相关的百度关键词,收集同期不同关键词的百度指数。采用 Spearman 相关分析关键词与流感病毒阳性率在先行、滞后 2 周时间范围内的时滞相关性,选择相关性最大的滞后期分别建立多元线性回归、差分自回归移动平均模型(auto-regressive moving average model, ARIMA)、支持向量机(support vector machine,SVM)、极端梯度提升算法(extreme gradient boosting,XGBoost)、随机森林五种模型,并选择 R^2、均方根误差(root mean squared error,RMSE)和平均绝对误差(mean absolute error,MAE)作为模型评价指标。**结果**　共筛选出 7 个关键词“流感、甲型流感、乙型流感、乙流、发烧、奥司他韦、达菲”纳入模型。其中 XGBoost 模型为最优模型,其模型的拟合 R^2、RMSE、MAE 分别为 0.923、3.848、2.434;模型的预测 R^2、RMSE、MAE 分别为 0.609、7.503、5.334。**结论**　基于百度指数构建流感预测模型具有一定预警性,能够为青海省实现科学精准的流感防控提供参考依据。

关键词:流感;百度指数;机器学习;预测
中图分类号:R181.3;R511.7　　**文献标志码**:A　　**文章编号**:1003-8507(2026)17-3099-06
DOI:10.20043/j.cnki.MPM.202603454

Comparison of machine learning models for influenza prediction in Qinghai Province based on Baidu Index

WANG Hui*, SHEN Rui, ZHANG Ming-hui, JIN Hai-xiao, WANG Zhao-cai, FENG Lu-zhao, XU Li-li
Department of Public Health, Qinghai University Medical College, Xining, Qinghai 810001, China

Abstract: Objective To analyze the correlation between influenza in Qinghai Province and Baidu Index, as well as the characteristics of the time lag changes, and establish an influenza prediction model applicable to Qinghai Province. **Method** Influenza virus positive rate monitoring data for Qinghai Province from 2015 to 2024 were obtained from the “China Influenza Monitoring Information System”, and Baidu keywords related to influenza were mined. The Baidu Index of different keywords during the same period was collected. Spearman correlation analysis was used to analyze the lag correlation between the keywords and the influenza virus positive rate within a lead and lag period of 2 weeks. The lag period with the highest correlation was selected to establish five models: multiple linear regression, auto-regressive integrated moving average model (ARIMA), support vector machine (SVM), extreme gradient boosting (XGBoost), and random forest (RF). R-square (R^2), root mean square error (RMSE), and mean absolute error (MAE) were selected as the evaluation indicators for the models. **Results** A total of 7 keywords “Influenza, Type A Influenza, Type B Influenza, Influenza B, Fever, Oseltamivir, Tamiflu” were selected and included in the model. Among them, the XGBoost model was the optimal model, with the fitting R^2, RMSE, and MAE of the model being 0.923, 3.848, and 2.434 respectively. The prediction R^2, RMSE, and MAE of the model were 0.609, 7.503, and 5.334 respectively. **Conclusion** The influenza prediction model constructed based on Baidu Index has certain early-warning capability and can provide a reference basis for Qinghai Province to achieve scientific and precise influenza prevention and control.

Keywords: Influenza; Baidu Index; Machine learning; Forecast

流行性感冒(简称流感),是由流感病毒感染引起的急性上呼吸道传染病,该病毒具有传染性强、传播速度快、人群普遍易感等特点[1-2]。我国每年流感相关呼吸道死亡病例有 8.4 ~ 9.2 万例[3],2011—2019 年流感造成的经济损失高达 1 060 亿元[4]。尽管我国建立了完善的流感监测系统,但常规监测信息需逐级报告和核实,因此获得的流感监测信息普遍存在 1 ~ 2 周的滞后[5]。

随着互联网的发展和普及,社交媒体上积累了大

基金项目:青海省昆仑英才·高层次卫生健康人才项目〔2025〕10 号
作者简介:汪晖(2000—),男,硕士在读,研究方向:传染病监测预警
通信作者:徐莉立,E-mail: qhcdcxull@126.com

量有关健康问题网络数据，这些信息数据具有较好的实时性，国外早期使用网络搜索数据改进流感监测系统，有效地解决了传统监测系统时间上的滞后问题[6-7]，同时也利用 Google、Twitter 等搜索数据开展了流感相关的预测研究[8-9]。百度是我国居民最常用的网络搜索平台，百度搜索数据能够实时反映公众对流感相关话题的关注度与健康需求。国内许多学者也在尝试利用百度指数进行疾病预测，包括流感的预测[10-11]。本研究利用青海省 2015—2024 年流感监测数据和百度搜索数据，基于不同模型算法构建适用于青海省流感预测模型并进行评价。

1　资料与方法

1.1　资料来源

1.1.1　流感监测数据　从“中国流感监测信息系统”获得 2015—2024 年青海省流感样病例（influenza-like illness, ILI）监测数据。依据《全国流感监测方案（2017 年版）》[12]规定，将发热（体温≥38℃）且伴有咳嗽或咽痛之一者确诊为 ILI，ILI 数占门、急诊就诊总数的百分比定义为 ILI%。青海省在 8 个市州共设立 20 所国家级流感哨点医院和 10 家网络实验室，每个流感监测哨点医院采集具有典型症状的 ILI 患者的鼻咽拭子样本，送往实验室进行病原学检测，流感病原阳性样本数占送检样本总数百分比定义为流感病毒阳性率。按照流行期（10 月—次年 3 月）每周采集 20 份，非流行期（4 月—9 月）每周采集 5 份，标本保存于 4℃的环境中，48 h 内送往网络实验室，实验室 3 个工作日内完成流感病毒鉴定。各监测年度中 90%以上的病例标本均能按上述要求送检，可获得较为准确的流感病原学监测信息。

1.1.2　百度指数数据　通过站长之家工具（https://data.chinaz.com/ci）进行关键词长尾词的挖掘，以“流感”为核心词汇，检索与流感高度相关的关键词。从百度指数官网（http://index.baidu.com/）获得 2015—2024 年青海省不同关键词每周“PC+ 移动”端的百度搜索指数。百度指数 = 特定关键词某段时间内的搜索量 / 百度搜索引擎总体搜索量 × 100。

1.2　预警方法

1.2.1　关键词时滞相关性分析　考虑流感从病毒感染到出现症状，具有一定的时间滞后性特征，关键词所反映的信息与流感发病之间的关联也可能存在时间上的错位。本研究选用 Spearman 相关分析关键词在先行、滞后 2 周时间范围内与流感病毒阳性率的相关性，选择不同关键词相关系数最高滞后期构建预测模型。

1.2.2　模型选择　将 2015 年第 1 周—2023 年第 52 周的数据作为训练集，2024 年第 1—52 周的数据作为测试集分别构建多元线性回归、差分自回归移动平均模型（auto-regressive moving average model, ARIMA）、随机森林、极端梯度提升算法(extreme gradient boosting,XGBoost)、支持向量机(support vector machine,SVM)、五个模型。

1.2.2.1　多元线性回归　使用 stats 包的 lm 函数建立多元线性回归模型。

1.2.2.2　ARIMAX　纳入 X 变量（百度指数）构建 ARIMAX(p,d,q)(P,D,Q)$_s$，单位根检验和 nsdiffs 函数确定 d 和 D，观察自相关函数图和偏自相关函数图确定 P 和 q，Q 和 P 分别取 1、2、3，S 为 52。

1.2.2.3　随机森林　调用 caret 包对随机森林模型 mtry 参数调优，创建调参网格，使 mtry 分别取值为 2、3 和 4，并采用 5 折交叉验证来评估不同参数设置下模型的性能表现。

1.2.2.4　极端梯度提升算法（extreme gradient boosting，XGBoost）　对于 XGBoost 模型参数，max_depth 设置为 6，eta 设置为 0.1，子采样比例和特征采样比例设置为 0.8，采用 5 折交叉验证结合早停机制确定最优迭代次数。

1.2.2.5　支持向量机（support vector machine，SVM）　使用 expand.grid 函数设置超参数调优网格，参数 sigma 取值为 0.01、0.1 和 1，C 取值为 0.1、1、10 和 100，利用 trainControl 函数配置 5 折交叉验证，确定 sigma 最佳参数。

1.2.3　模型评价指标　为了能够客观准确地比较不同模型的拟合预测效果，本次研究中使用 R 方（R^2）、均方根误差（root mean squared error，RMSE）和平均绝对误差（mean absolute error，MAE）作为模型性能的评估指标。其计算公式为：

$$R^2=1-\frac{\sum_i(\hat{y}^{(i)}-y^{(i)})^2}{\sum_i(\bar{y}^{(i)}-y^{(i)})^2}$$　公式（1）

$$RMSE=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}$$　公式（2）

$$MAE=\frac{1}{n}\sum_{i=1}^{n}\left|y_i-\hat{y}_i\right|$$　公式（3）

其中，$\hat{y}_i$ 表示预测值，y_i 表示真实值；n 代表样本个数。

1.3　质量控制

按《全国流感监测方案（2017 年版）》和《全国流感监测质量评估方案（2017 年版）》要求完成 ILI 采样、保存、运送、实验室检测等环节，相关信息及时录入“中国流感监测系统”，并由专业人员审核，确保流感病原学数据的完整性和准确性；百度指数数据通过对比官方数据验证其数据的可靠性。缺失值采用邻近值补齐法补齐，最后，青海省 2015—2024 年流感病毒阳性率数据和百度指数数据统一按

周整理，共 522 周。

1.4　统计分析　采用 Excel 2010 进行数据整理和描述分析，运用 SPSS 25.0 进行 Spearman 相关性分析，$P<0.01$ 具有统计学意义，使用 R 4.3.2 软件中的“stats、arima、caret、xgboost”程序包完成模型的构建。

2　结　果

2.1　流感病毒阳性率与百度指数变化趋势　2015—2024 年青海省流感病毒阳性率在时间上呈现明显的季节性和暴发流行阶段。围绕流感的称谓、症状和治疗方面获得 8 个搜索量高的关键词，分别为流感、甲型流感、乙型流感、乙流、发烧、奥司他韦、达菲、咽痛，在 2017—2020 和 2022—2024 年两个时间段内，三类关键词的搜索指数与流感病毒阳性率变化趋势更加吻合，见图 1。其中，与治疗相关的关键词（奥司他韦、达菲）搜索指数变化趋势与流感病毒阳性率的波动最贴合，尤其是阳性率出现极峰时；与称谓相关的关键词（流感、甲型流感、乙流、乙型流感）仅在流感病毒阳性率出现高峰时同步增长；但有关症状的关键词（发烧、咽痛）与流感病毒阳性率变化趋势的一致性相对较弱。

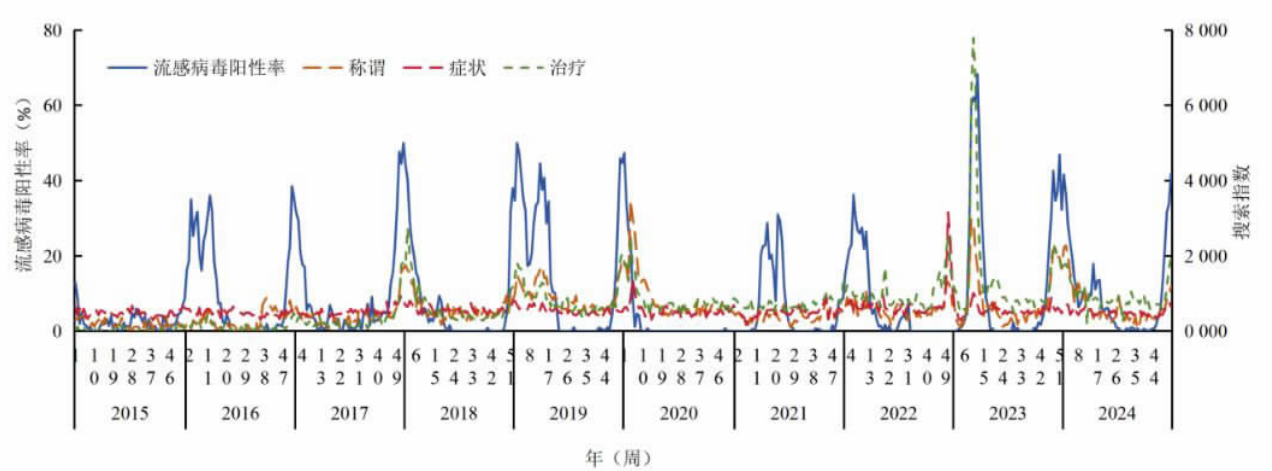

图 1　流感病毒阳性率与百度关键词搜索指数变化趋势

Figure 1　The positive rate of the influenza virus and the trend of changes in Baidu keyword search index

2.2　时滞相关性分析　根据不同滞后周期，将不同关键词分为先行、同步及滞后关键词。分别计算关键词不同时期与流感病毒阳性率的相关系数，见表 1。时滞相关性分析结果显示，“乙型流感、达菲”两个关键词在先行 2 周时相关系数达到最大；“乙流、发烧、奥司他韦”三个关键词在先行 1 周时相关系数最大；关键词“甲型流感”在同步时相关系数最大；关键词“流感”则在滞后 1 周时相关系数最大；将上述 7 个关键词纳入模型，进行建模预测。

表 1　流感病毒阳性率与百度关键词的相关系数

Table 1　The correlation coefficient between the positive rate of the influenza virus and the keywords on Baidu

关键词		先行 2 周	先行 1 周	同步	滞后 1 周	滞后 2 周
流感	r	0.082	0.098	0.086	0.113*	0.096
	P 值	0.058	0.025	0.048	0.009	0.028
甲型流感	r	0.288	0.297	0.309*	0.309	0.262
	P 值	<0.001	<0.001	<0.001	<0.001	<0.001
乙型流感	r	0.448*	0.421	0.392	0.348	0.312
	P 值	<0.001	<0.001	<0.001	<0.001	<0.001
乙流	r	0.405	0.373*	0.343	0.306	0.263
	P 值	<0.001	<0.001	<0.001	<0.001	<0.001
发烧	r	0.373	0.363*	0.358	0.335	0.299
	P 值	<0.001	<0.001	<0.001	<0.001	<0.001
奥司他韦	r	0.253	0.260*	0.247	0.238	0.195
	P 值	<0.001	<0.001	<0.001	<0.001	<0.001
达菲	r	0.196*	0.177	0.165	0.153	0.138
	P 值	<0.001	<0.001	<0.001	<0.001	<0.001
咽痛	r	−0.041	−0.043	−0.082	−0.104	−0.103
	P 值	0.354	0.330	0.063	0.018	0.019

注：* 最大相关系数。

2.3　模型预测结果　依次计算出 5 个模型的 R^2、RMSE 和 MAE 指标，结果显示，XGBoost 模型在训练集和测试集均表现最优，在拟合训练方面，R^2 为 0.923，RMSE 和 MAE 分别为 3.848、2.434。从预测效果来看，R^2 达到最大为 0.609，RMSE 和 MAE 最小为 7.502、5.334。不同模型的评价指标结果见表 2。

2.4　模型比较　通过对比不同模型的预测值与实际值发现，不同模型在预测流感病毒阳性率方面存在差异。其中，SVM 和 ARIMAX 模型预测效果较差，尤其在流感流行高峰期，SVM 模型的预测值与实际值偏离较大。相比之下，XGBoost 模型展现了更好的预测性能，不论是在流感病毒阳性率的高峰期还是低谷期均能够更好地捕捉到时间序列上的变化特征，表现出较为理想的预测效果。XGBoost 模型是本研究中预测准确度最优模型，可以精准地预测流感病毒阳性率的变化情况。不同模型对青海省 2024 年 1—52 周流感病毒阳性率的预测效果见图 2。

表 2　不同模型评价对比

Table 2　Comparison of different models' evaluations

模型		R^2	RMSE	MAE
SVM	训练集	0.767	6.537	3.411
	测试集	0.466	11.319	7.503
ARIMAX	训练集	0.907	4.039	2.575
	测试集	0.575	8.389	7.138
随机森林	训练集	0.916	4.160	2.759
	测试集	0.590	7.397	3.968
多元线性回归	训练集	0.548	8.865	6.057
	测试集	0.595	7.219	5.357
XGBoost	训练集	0.923	3.848	2.434
	测试集	0.609	7.503	5.334

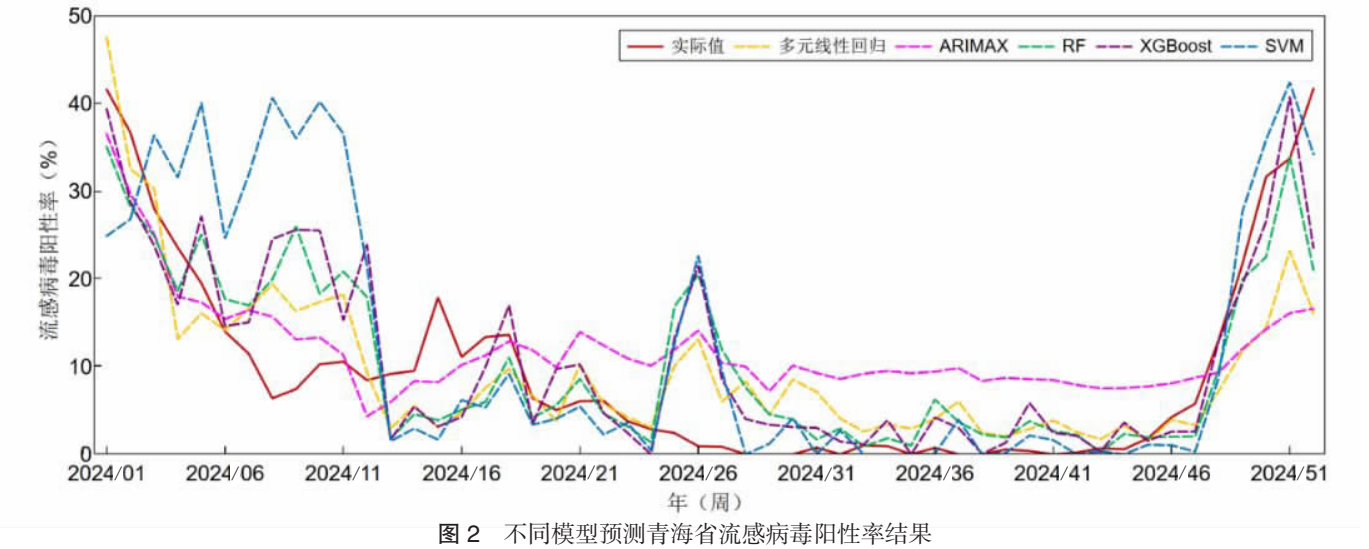

图 2　不同模型预测青海省流感病毒阳性率结果

Figure 2　Prediction results of the influenza virus positivity rate in Qinghai Province using different models

2.5　敏感性分析　为验证模型内部稳健性并检验模型对变量的敏感程度，本研究使用流感的称谓、治疗和症状三个粗维度变量构建建敏感性分析模型，其中症状与治疗在先行 2 周时相关系数最大，称谓在先行 1 周时相关系数最大。选择最优模型 XGBoost 建立敏感性分析模型，其预测 R^2 为 0.563，RMSE 和 MAE 分别为 7.805 和 6.075。相比纳入 7 个具体关键词的 XGBoost 模型，其预测预测精度降低 4.6%，表明输入变量的变化会直接影响模型的预测能力，模型对输入变量具有较好的敏感性。

3　讨　论

由于传统流感监测体系和时间序列的预测预警普遍存在及时性差和精准度低的问题，难以满足疾控部门对疫情的快速决策和处置需要[13]。而互联网搜索行为能够直观反映公众对健康相关信息的关注与需求，可以为传染病监测预警提供新型数据源。国内外学者证实利用新闻报道、社交媒体等非正式来源数据可以对流感进行更好地实时监测[11]，自 2009 年 Ginsberg[6]团队使用网络搜索成功预测流感，许多学者利用 Google[14]、Wikipedia[15]等搜索数据开展了流感预测研究，基于互联网搜索数据预测疾病也成为新的聚焦点。在我国，百度搜索引擎是国民长期使用的网络平台，百度搜索指数具有良好的网络搜索信息代表性[10]。截至 2025 年，我国网民数量达到 11.23 亿人[16]，这也意味着我国拥有足够的网络搜索数据，能够开展基于网络搜索的疾病预测研究[17]。

本研究通过挖掘与流感相关的百度搜索关键词，围绕流感的称谓、症状和治疗，初步选择搜索量较高的 8 个关键词纳入研究，通过分析不同关键词与流感病毒阳性率的时滞相关性，选择“流感、甲型流感、乙型流感、乙流、发烧、奥司他韦、达菲”关键词相关性最大的滞后期建立流感病毒阳性率预测模型。关键词“乙型流感”是对模型影响最大的特征，模型性能由高

到低依次为：XGBoost、多元线性回归、随机森林、ARIMAX、SVM，XGBoost 模型能够更好地预测青海省流感病毒阳性率变化趋势。

在时间维度上对比不同类别关键词搜索指数与流感病毒阳性率之间的变化趋势，结果发现，与流感治疗相关的关键词“奥司他韦、达菲”的搜索指数变化趋势与流感病毒阳性率具有较高的一致性，此现象可能与“奥司他韦、达菲”均为抗病毒药物有关，临床医生在明确患者感染流感病毒后开具此类药物，患者出于自身健康关注，通过网络搜索了解相关药物服用方法、副作用等相关信息[18]，使得治疗相关关键词的百度指数变化趋势与流感病毒阳性率出现高度一致。同时，这也提示抗病毒药物监测记录在流感监测预警中具备较高的价值，可将其作为流感多渠道监测的社会感知敏感指标。然而，流感病毒阳性率与“发烧、咽痛”的搜索指数变化趋势一致性相对较弱，造成此现象的原因可能是其他呼吸道病原体的感染，例如肺炎衣原体和腺病毒，同样会引发“发烧、咽痛”等症状，导致相关关键词的百度指数在非流感高峰期上升，弱化了这类关键词作为流感的特异性信号[19]。

时滞相关性分析结果发现，“甲型流感、乙型流感、乙流、发烧”等关键词的最大相关系数均在 0.3 ~ 0.6 之间，与清远市研究结果相似[20]。但在探索百度搜索关键词与流感病毒阳性率相关性时，不仅需要分析互相关强度，时间维度上的对应性也很重要，具有领先性或同步性的关键词具备较高的预测价值[21]。本研究中关键词“乙型流感、达菲”和“乙流、发烧、奥司他韦”与流感病毒阳性率之间分别存在领先 2 周和 1 周的相关性，这些提示特定关键词可以作为青海省流感监测预警的先行指标，提供早期预警信息。此外，本研究发现不同关键词的百度指数与流感病毒阳性率之间的相关性强弱会随着不同滞后期发生改变，这表明在构建青海省流感监测预警体系时，不仅要拓展关键词的选择范围，关注流感发生发展的潜在信号，还需要更加细致地分析各关键词在不同滞后期的相关性强度变化规律，以最强相关性为核心构建流感监测预警体系，才能及时、准确地预测我省流感疫情的实际变化情况，从而在流感防控工作中占据主动。

在基于百度指数数据构建疾病预测模型的研究中，李云兰[22]的中国脑卒中患者预测研究和黄善君[21]等人的吉林省水痘预测研究中，均发现 XGBoost 模型具有较好的预测性能。本研究中 XGBoost 模型为最优模型，XGBoost 模型的预测 R^2 为 0.609，RMSE 为 7.503，MAE 为 5.334，其预测精度较 SVM 模型提升了约 14.3%，较 ARIMAX 模型提升了约 3.4%。与薛红新等人[11]结合百度指数建立中国南方地区流感预测模型的研究结论相似，进一验证了 XGBoost 模型在基于百度指数的疾病预测研究中的优越性，这表明 XGBoost 模型在捕捉非线性关系、滞后效应等复杂的数据特征方面更具有优势。敏感性分析结果显示，XGBoost 模型在变量调整后，预测性能虽然降低，但仍维持在可接受水平，表明 XGBoost 模型也具备较好的敏感性和稳健性。与此同时，本研究的不同之处在于聚焦青藏高原地区青海省，其气候条件、经济条件及卫生资源配置与其他研究区域存在显著差异，流感流行特征与公众搜索行为模式存在不同。其次，相较于既往研究中常用的 ILI 和 ILI%，流感病毒阳性率能够更直接、精准地反映流感活动水平。上述两方面使本研究的研究结论更具区域针对性，后续可以考虑在本研究基础上，探索更加精准反映青海省流感活动水平的数据，选用更复杂、能够处理高维数据的机器学习或深度学习模型开展预测预警工作，进一步提升预测的精准度和可靠性。

在公共卫生防控不断迈向智能化、精准化的当下，本研究聚焦青海省流感疫情，其研究结果对青海省流感疫情防控工作具有参考价值。百度指数和机器学习模型可作为智慧化多触点流感监测预警体系的重要组成部分，为青海省流感的实时监测和早期精准预测提供有力支撑。我省疾控部门可以依据模型预测结果及时或更早地采取相应的防控措施，例如扩大监测范围、加强易感人群疫苗接种、提醒公众加强自我防护等，以有效遏制流感疫情的蔓延[13]。此外，本研究所使用的研究方法与预测模型，可为我省手足口病、感染性腹泻等其他常见传染病监测预警提供方法学的借鉴与参考。

本研究存在不足之处，百度搜索平台尽管在我国互联网市场占据较大市场份额，但是近年来网络新兴媒体的发展使得网民在目的信息搜索时拥有多元选择空间，可以根据个人喜好选择不同的网络平台，这在一定程度上削弱了百度搜索引擎的代表性，后续可以考虑整合多个主流网络平台的搜索数据进行分析。

综合上述研究结果，基于百度指数可以构建青海省流感预测模型，并且展现出良好的预测效能，能够更加精准地捕捉流感流行趋势，可以为建立青海省流感的监测预警体系提供参考。未来研究中，可以考虑纳入多源数据，采用深度学习、神经网络等算法针对青海省不同流感病毒亚型的变化趋势进行预测研究，为青海省公共卫生预测预警提供更加科学可靠的决策支持。

利益冲突声明　本研究不存在任何利益冲突

参考文献

[1] 姜艳艳,田蒙蒙,申亚琦,等. 2017-2023 年山东省潍坊市流行性感冒病毒病原学监测结果分析[J]. 预防医学论坛,2024,30(7):528-531, 536.
Jiang YY, Tian MM, Shen YQ, et al. Analysis of the epidemiological monitoring results of influenza viruses in Weifang City, Shandong Province from 2017 to 2023 [J]. Chinese Journal of Preventive Medicine, 2024, 30(7): 528-531, 536.(In Chinese)

[2] 李静,赵艺明,王猛,等. 2017-2023 年上海市金山区流感样病例监测结果分析[J]. 华南预防医学,2025,51(8):903-907.
Li J, Zhao YM, Wang M, et al. Analysis of influenza-like case monitoring results in Jinshan District, Shanghai from 2017 to 2023 [J]. South China Journal of Preventive Medicine, 2025, 51 (8): 903-907.(In Chinese)

[3] Diamond C, Gong H, Sun FY, et al. Regional-based within-year seasonal variations in influenza-related health outcomes across mainland China: a systematic review and spatio-temporal analysis[J]. BMC Medicine, 2022, 20(1): 58.

[4] 安惠泽. 中国流感发病率时空趋势、影响因素及预测方法研究[D]. 太原:山西财经大学,2024.
An HZ. Research on the spatial-temporal trend, influencing factors and prediction methods of influenza incidence in China[D]. Taiyuan: Shanxi University of Finance and Economics, 2024.(In Chinese)

[5] Shen R, Xu XY, Yang L, et al. Real‐time digital prescriptions unlock influenza dynamics: evidence from 21 million transactions[J]. NPJ Digital Medicine, 2026, 9(1): 315.

[6] Ginsberg J, Mohebbi MH, Patel RS, et al. Detecting influenza epidemics using search engine query data [J]. Nature, 2009, 457 (7232): 1012-1014.

[7] Mcclymont H, Lambert SB, Barr I, et al. Internet-based Surveillance Systems and Infectious Diseases Prediction: An Updated Review of the Last 10 Years and Lessons from the COVID-19 Pandemic [J]. Journal of Epidemiology and Global Health, 2024, 14(3): 645-657.

[8] Liscano Y, Anillo Arrieta LA, Montenegro JF, et al. Early warning of infectious disease outbreaks using social media and digital data: a scoping review [J]. International Journal of Environmental Research and Public Health, 2025, 22(7): 1104.

[9] Martín-Corral M, David, García-Herranz G, et al. Social media sensors as early signals of influenza outbreaks at scale [J]. EPJ Data Science, 2024, 13(1): 239.

[10] Dai SF, Han LT. Influenza surveillance with Baidu index and attention-based long short-term memory model[J]. PLOS One, 2023, 18(1): e0280834.

[11] 薛红新,刘绕星,况立群,等.基于百度指数与机器学习方法的流感样病例预测[J]. 中国预防医学杂志,2023,24(8):788-794.
Xue HX, Liu RX, Kuang LQ, et al. Influenza-like illness prediction based on Baidu index and machine learning method [J]. China Preventive Medicine, 2023, 24(8): 788-794.(In Chinese)

[12] 中国疾病预防控制中心. 全国流感监测技术指南(2017 年版)[EB/OL]. [2026-07-07]. http://www.chinaivdc.cn/cnic/fascc/201802/t20180202_158592.htm.
Chinese Center for Disease Control and Prevention. National influenza surveillance technical guidelines (2017 Edition)[EB/OL]. [2026-07-07]. http://www.chinaivdc.cn/cnic/fascc/201802/t20180202_158592.htm.(In Chinese)

[13] 于树悦,梁焱榆,陈立凌,等. 基于机器学习整合气象、互联网和监测数据的流感预测模型构建与评价 [J]. 现代医学,2024,52(11):1695-1702.
Yu SY, Liang YY, Chen LL, et al. Construction and evaluation of influenza prediction models based on the integration of meteorological, Internet,and surveillance data using machine learning [J]. Modern Medical Journal, 2024, 52(11): 1695-1702.(In Chinese)

[14] Nsoesie EO, Oladeji O, Abah ASA, et al. Forecasting influenza-like illness trends in Cameroon using Google Search Data [J]. Scientific Reports, 2021, 11(1): 6713.

[15] De toni G, Consonni C, Montresor A. A general method for estimating the prevalence of influenza-like-symptoms with Wikipedia data[J]. PLOS One, 2021, 16(8): e0256858.

[16] 佚名. 11.23 亿网民如何共享数字发展成果?-- 透视第 56 次《中国互联网络发展状况统计报告[J].信息系统工程,2025,(8):2.
Anonym. How do 1.123 billion internet users share the fruits of digital development? —— Insights from the 56th "Statistical Report on the Development of China's Internet" [J]. Information System Engineering, 2025, (8): 2.(In Chinese)

[17] 李进,魏艳龙,薛红新. 基于随机森林模型与 SHAP 算法的流感样病例影响因素分析研究 [J]. 信息技术与信息化,2024,(2):3-6.
Li J, Wei YL, Xue HX. Analysis of influencing factors of influenza-like cases based on random forest model and SHAP algorithm [J]. Information Technology & Informatization, 2024, (2): 3-6.(In Chinese)

[18] 李乐,周子豪,吴群红,等. 流感数据与特定关键词相关性分析及预测[J]. 中国公共卫生,2021,37(12):1813-1818.
Li L, Zhou ZH, Wu QH, et al. Association of specific keywords index in Baidu website with influenza monitoring data during 2012-2020 and potential use of the index for influenza epidemic prediction[J]. Chinese Journal of Public Health, 2021, 37(12): 1813-1818.(In Chinese)

[19] 加吾兰·祖农,苏承,杨杰,等. 基于特定关键词的百度指数与天津市流感、新冠监测指标相关性分析 [J]. 首都公共卫生,2025,19(2):75-80.
Jia Wu Lan - Zu Nong, Su C, Yang J, et al. Correlation analysis of specific keywords' Baidu index with influenza and COVID-19 surveillance indicators in Tianjin [J]. Capital Journal of Public Health, 2025, 19(2): 75-80.(In Chinese)

[20] 杜玉忠,范秀红,卢文涛,等. 百度关键词在清远市流行性感冒发病预测中的应用及效果评价 [J]. 中国热带医学,2018,18(8):792-794.
Du YZ, Fan XH, Lu WT, et al. Application and effect evaluation of Baidu keywords in predicting the incidence of influenza in Qingyuan City[J]. Chinese Tropical Medicine, 2018, 18(8): 792-794.(In Chinese)

[21] 黄善君. 2014-2023 年吉林省水痘流行特征及基于百度指数的预测研究[J]. 长春:吉林大学,2025.
Huang SJ. Research on the epidemiological characteristics of varicella in Jilin Province from 2014 to 2023 and prediction based on Baidu index[J]. Changchun: Jilin University, 2025.(In Chinese)

[22] 李云兰. 基于百度指数的中国脑卒中患病预测模型研究[D]. 衡阳:南华大学,2024.
Li YL. Research on prediction model of stroke incidence in China based on Baidu index [D]. Hengyang: University of South China, 2024.(In Chinese)

收稿日期:2026-03-24