基于互联网数据的传染病预测模型研究进展*

何琪乐[1] 张瑾瑶[1] 吴卓存[1] 杨予青[1] 赵 伟[2] 胡红濮[1]

([1] 中国医学科学院/北京协和医学院医学信息研究所 北京 100020 [2] 北京市垂杨柳医院 北京 100022)

〔摘要〕 **目的/意义** 系统梳理基于互联网数据的传染病预测模型相关研究，助力实现传染病监测关口前移，为构建传染病智慧化立体防治体系提供参考。**方法/过程** 对 Web of Science 核心数据库和中国知网收录的近 20 年基于互联网数据的传染病监测预警研究发展历程及研究方向进行梳理，分析当前主要问题与挑战，总结常见预测模型及其优化方向。**结果/结论** 互联网传染病监测研究呈监测疾病多样化、数据来源精细化和专业化等趋势。由于互联网数据的复杂性和不确定性，现有模型大多仅适用于短时或实时预测。通过构建组合模型、加强多源数据融合、完善关键词与影响因素选择等方式，可进一步优化模型，加强拟合效果和预测能力。

〔关键词〕 传染病监测预警；流行病情报学；预测模型；搜索引擎；互联网

〔中图分类号〕 R-058 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673-6036.2024.02.006

Research Progress of the Infectious Disease Prediction Models Based on Internet Data

HE Qile[1], *ZHANG Jinyao*[1], *WU Zhuocun*[1], *YANG Yuqing*[1], *ZHAO Wei*[2], *HU Hongpu*[1]

[1] *Institute of Medical Information, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing 100020, China*; [2] *Beijing Chuiyangliu Hospital, Beijing 100022, China*

〔**Abstract**〕 **Purpose/Significance** The paper systematically reviews relevant research on infectious disease prediction models based on internet data, helps to realize the advancement of infectious disease surveillance, and provides references for the construction of intelligent three-dimensional prevention and treatment system of infectious diseases. **Method/Process** The development history and research direction of infectious disease surveillance and early warning based on internet data collected in the core database of Web of Science and CNKI in the past 20 years are reviewed, major existing problems and challenges are analyzed, and common prediction models and their optimization directions are summarized. **Result/Conclusion** The study on internet infectious disease surveillance shows the trend of diversification of monitoring diseases, refinement and specialization of data sources. Due to the complexity and uncertainty of internet data, most of the existing models are only suitable for short-term or real-time prediction. By constructing a combination model, strengthening multi-source data fusion, improving the selection of keywords and influencing factors, the model can be further optimized and the fitting effect and prediction capacity can be strengthened.

〔**Keywords**〕 infectious disease surveillance and early warning; epidemic intelligence; prediction model; search engine; internet

〔修回日期〕 2024-01-16

〔作者简介〕 何琪乐，硕士研究生，发表论文 5 篇；通信作者：胡红濮，研究员，博士生导师。

〔基金项目〕 国家社会科学基金重点项目（项目编号：22AZD089）；国家社会科学基金重大项目（项目编号：22&ZD141）；中国医学科学院医学与健康科技创新工程（项目编号：2022-I2M-1-019）。

1 引言

对传染病进行监测预警是控制其传播的重要手段。传统传染病监测主要依靠各级医疗机构、疾控中心和监测哨点医院等构成的监测网，虽然准确性高但监测速度通常滞后于传播速度，且应对新发传染病时数据来源较少。基于 Web of Science 核心期刊数据库和中国知网，以 TS = ((epidemic AND (monitoring OR surveillance OR forecast OR predict OR warning) AND (internet OR "search engine" OR "social media")) OR "epidemic intelligence") 和 (主题 = (传染病 OR 流行病) AND (监测 OR 预测 OR 预警) AND ("搜索引擎" OR "大数据" OR "互联网")) OR (主题 = "信息流行病学") 为主题词检索式，对 2001—2022 年发表的基于互联网数据的传染病预测相关文献进行检索，查得英文文献 864 篇，中文文献 162 篇。分析检索结果发现，互联网数据可用于传染病监测预警已成为研究共识[1]，且相关论文发表数量趋势，见图 1。在既往研究基础上，本研究从基于互联网数据的传染病监测预警研究发展历程、应用场景、常见预测模型、主要问题与挑战、发展趋势等方面进行探讨，旨在为进一步建立基于大数据、人工智能等新技术的智慧公共卫生应急管理模式提供参考依据。

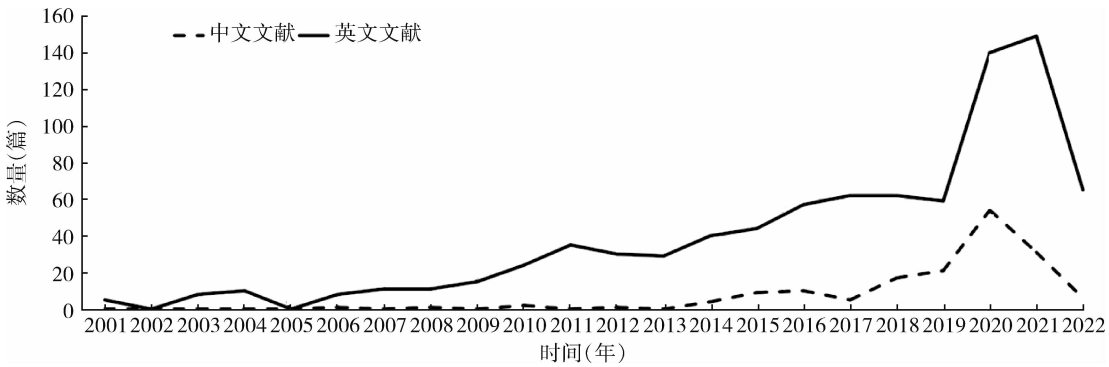

图 1 2001—2022 年国内外基于互联网数据的传染病预测相关论文发表数量趋势

2 基于互联网数据的传染病监测预警应用场景

互联网传染病监测数据源可分为搜索引擎结构化数据和社交媒体文本数据。基于搜索引擎数据的研究主要开展基于关键词检索指数和传染病上报数据的时差相关性分析，构建不同滞后期的复合关键词及搜索指数[2-3]。文本数据主要来源于推特、微博等社交媒体。在前期文献检索的基础上，补充结合文献计量主题词相关结果，统计 2001—2022 年国内外热点疾病相关论文年发表数量，共计 272 篇，分类绘制气泡图，圆圈直径大小反映论文数量的多少，见图 2。分析可得，流感、肺结核、艾滋病、登革热、埃博拉、寨卡、乙型病毒性肝炎（乙肝）等疾病为研究热点。

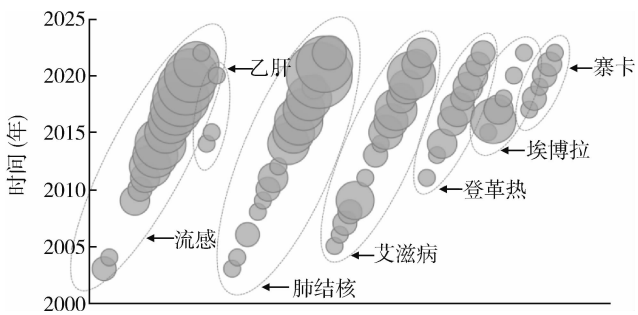

图 2 2001—2022 年国内外各传染病相关论文发表数量及趋势

流感是最早将网络搜索数据纳入监测系统的传染病，以谷歌流感趋势最具代表性。但由于其准确性会受到用户搜索行为、传染病季节性等因素影响，其预测的流感发病率高于美国疾控中心的实际报告值[4]。Luo Y 等[5]融合多来源搜索数据预测 2009 年甲型 H1N1 流感的流行规模，发现较单一搜

索引擎的拟合效果更优；Mauricio S 等[6]以医学专业网站 UpToDate 及医学专业词汇作为数据源和关键词预测流感，发现专业网站可靠性更强。搜索数据融合地理位置及环境因素可进一步获得较理想的监测效果。Gluskin R T 等[7]提出谷歌登革热趋势，发现在高流行地区和登革热传播适宜气候中准确性更高；Zhou X 等[8]分别拟合并比较动态模型和线性回归模型在不同地域层次上的肺结核监测能力；唐家博[9]以手足口病为监测预警对象，对互联网和气象数据进行挖掘。

3 基于互联网的传染病预测模型种类与比较

3.1 简单回归预测模型

多元线性回归是常见的简单回归模型之一。Bodnar T 等[10]将其用于流感监测，发现可以通过为每个检索关键词分配不同权重减少干扰词汇产生的噪声。但解释变量之间可能有多重共线性，且向后剔除变量时会减少原数据信息。

3.2 时序预测模型

3.2.1 统计学模型 常用于互联网数据传染病监测的统计学模型包括自回归移动平均（auto - regressive integrated moving average，ARIMA）模型和动态线性模型（dynamic linear model，DLM）。ARIMA 可将非平稳的时间序列平稳化，将因变量对其滞后值和随机误差项的现值和滞后值进行回归，有效提取具有季节性和趋势性的数据中的线性信息，但对非线性、无规律、波动大的数据和长期预测效果较差[11]。DLM 是一种高斯线性状态空间模型，可用于对非平稳时间序列进行建模，包括测量方程和状态方程。测量方程可以根据某时刻的参数向量描述此时对应的因变量，状态方程可以建立该时刻的参数向量和下一时刻的参数向量之间的联系，从而进行预测[12]。

3.2.2 传统机器学习模型 （1）随机森林（random forest，RF）。是对多个弱分类器进行组合的有监督学习，具有较高准确性和泛化性能[13]。Amin S 等[14]通过分析 2017—2019 年推特中关于疾病情绪的社交媒体文本，监测登革热和流感，并发现 RF 在提高准确度、精度、召回率等方面均优于比较模型。张金宇[15]以 2017—2019 年登革热流行情况为研究对象，发现 RF 预测效果整体较好，但不足以预测发病高峰。这可能是由于 RF 虽然能更好地削弱异常值对结果的影响，但导致差异度小的正确决策被淹没。（2）极端梯度提升（eXtreme gradient boosting，XGBoost）。是一种基于决策树的提升算法，使用多个分类树和回归树来学习输入变量和结果之间的非线性和复杂关系，可以更灵活地调整更多参数，整体上寻求最优解，在一定程度上避免过度拟合[16]。Meng D 等[17]针对手足口病建立了 RF 和 XGBoost 预测模型，发现从整体来看，XGBoost 较 RF 模型具有更好的预测能力。（3）支持向量机回归（support vector regression，SVR）。特点是通过非灵敏损失函数测量拟合优度，而非使用常规的二次损失函数（均方差）。Aramaki E 等[18]在进行流感相关推特文本分析时发现，SVR 具有最高精度和最短训练时间。但黄泽颖[19]发现多元线性回归模型相较于 SVR 能更好地拟合 2013—2018 年 H7N9 亚型禽流感新增病例数且预测精度更高。

3.2.3 深度学习模型 深度学习是机器学习领域中的新方向，其概念源于人工神经网络。人工神经网络模型擅长拟合复杂函数，形成非线性映射关系并行处理海量信息[20]。（1）BP 神经网络（back propagation neural networks，BP）。是一种广泛使用的神经网络模型，可以学习和存储大量无需用数学方程准确描述的输入 - 输出映射关系[21]。王若佳[22]使用 BP 模型，通过融合百度指数预测流感暴发。从拟合结果看，BP 神经网络的拟合效果比 SVR 更好，但拟合效果不等同于预测精度。此外，BP 模型很难引进时间维度，仅能使用当期搜索信息估计当期流感状况，故被称为临近预警模型。（2）广义回归神经网络（generalized regression neural network，GRNN）模型是一种基于数理统计的径向基函数网络，可以任意精度逼近非线性函数，解决了 BP 神经网络局部最优的问题。GRNN 的非线性映射能力和学习速度很强，且结构简单、收敛速度快，在传染病预测中得到广泛应用[23]。杨德志[24]建立

GRNN 模型和 BP 神经网络模型，发现 GRNN 的拟合和预测效果更好。（3）长短期记忆神经网络（long short - term memory，LSTM）模型是一种特殊的递归神经网络，可预测长时间滞后的时间序列，处理非线性成分并进行误差校正[25]。黄鹏[26]发现 LSTM 模型相较于 ARIMA 模型更适合用于乙类传染病预测研究；Parwez M A 等[27]使用推特活动即时预测当周发病率，证实了 LSTM 模型在预测误差最小情况下的有效性。

3.3 模型比较

总结既往研究发现，常见模型大多考察数据间的线性关系，非线性模型涉及人工神经网络常用模型，见表 1。由于搜索数据与真实数据之间关系的复杂性和较强的不确定性，在选择建模时应重点关注非线性模型，以获得更好的拟合效果和预测能力。此外，大多数模型仅适用于短时或实时预测，实现长时间段的预测较困难。

表 1 基于互联网数据的传染病预测常见模型比较

模型类型	适用性	优势	劣势
自回归移动平均（ARIMA）模型	能有效提取具有季节性与趋势性的时间序列中的线性信息，较好预测出时间序列的自相关性和季节性	结构简单、易于实现，预测精度较高	理想假设较多、要求严格，对非线性数据、无规律、波动大的数据进行长期预测时，效果欠佳
动态线性模型（DLM）	可用于对非平稳时间序列进行建模	更容易应用于不同类型的时间序列，且在新数据可用时不需要新的识别和建模周期	大型数据集计算成本高，对主观假设敏感，且存在过拟合风险
随机森林（RF）	一系列树模型的集合，可解决分类问题及回归问题	精确度和泛化性能较高，不易陷入过拟合，抗噪能力强	对发病高峰预测不足
极端梯度提升（XGBoost）	使用多个分类和回归树，以 Boost 集成的方式学习输入变量和结果之间的非线性关系	处理回归问题精度高，能充分利用多变量的潜在特征	处理趋势性不明显的传染病时表现不突出，可解释性较差
支持向量机回归（SVR）	以支持向量机作为数据挖掘方法处理时间序列分析问题	对异常值具有鲁棒性，决策模型可以轻松更新，具有出色的泛化能力和预测精度	在每个数据点的特征数量超过训练数据样本数量时表现不佳；当数据集有更多噪声时会出现目标类重叠
BP 神经网络模型	拟合复杂函数，形成非线性映射关系并行处理海量信息	无需用数学方程描述映射关系	难以引进时间维度，仅能预测当期状况，且易陷入局部最优
广义回归神经网络（GRNN）模型	是建立在数理统计基础上的径向基函数网络，理论基础是非线性回归分析，适用于小样本场景	非线性映射能力和学习速度强，网络结构简单，收敛速度快	空间复杂度高，测试样本全部的训练样本都要参与计算
长短期记忆神经网络（LSTM）模型	是一种特殊的递归神经网络，适用于自然语言处理、长期依赖关系和时序模式	善于处理和预测长时间滞后的时间序列，能处理非线性成分并进行误差校正，较好地处理多变量问题	复杂性高，训练和推理速度相对较慢，调参困难，预测结果有时难以解释

3.4 主要问题与挑战

虽然利用互联网信息进行传染病监测具有实时快速、数据源丰富、自动化程度高等优势，但仍存在很多不足。一是目前国内算法模型创新和疾病种类相对较少，多数研究仅使用 2 ~ 3 种模型预测方法，在关键词选择及变量合成方面也偏主观；从预测时间跨度来看，大多数模型仅适用于短时间预测。二是国内研究数据来源较单一，且存在较多混杂因素。互联网搜索行为可能会受到媒体报道、传染病季节性、互联网用户数量、文化差异、语言等因素影响，因此，基于搜索引擎的传染病预测仅能

体现相关性，无法完全替代传统监测[28]。

4 模型优化与发展趋势

4.1 构建组合模型

为弥补上述不足，可采用构建组合模型的方式提高预测精度。Su K 等[29]将季节性 ARIMA 模型和 XGBoost 模型相结合，构建具有自适应权重调整机制的 SAAIM 模型；赖晓莹等[30]构建 ARIMA－LSTM－XGBoost 加权组合模型，在预测精度上有较大提升；魏麟等[23]提出 CEEMD－GRNN 组合模型，精度更高、稳定性更强。

4.2 多源数据融合与加强

融合多源数据，包括各搜索引擎数据、社交网络数据，以及其他来源如智能穿戴设备数据、气象数据等，可提高模型的鲁棒性和泛化能力。Su K 等[29]收集重庆市流感样疾病历史百分比、气象数据、百度搜索指数和新浪微博数据等多源数据进行预测；Anwar M 等[31]同时使用谷歌和推特数据以提高模型准确率。

4.3 关键词与影响因素选择优化

关键词选择可能存在主观判断带来的局限性，因此改进选词方法对今后研究具有重要意义，应不断扩展可选词的范围、提高相关性和特异性。此外，地理位置因素（当地人口规模、生活习惯、互联网普及率等）对相关性影响较大，将来可在不同行政区的不同水平进行分层分析，以提供更有针对性的建议。

5 结语

传染病影响范围广泛，处置不及时可能造成严重损失。利用互联网数据对流行性疾病进行监测具有重要现实意义。由于其快速方便且成本低廉，在针对准确率进一步优化后，可作为传统监测网络的重要补充，辅助各地区疾控中心分析传染病的流行特征，从而制订相关防治策略和应急措施。

利益声明：所有作者均声明不存在利益冲突。

参考文献

1 MILINOVICH J G, WILLIAMS M G, CLEMENTS A C A, et al. Internet－based surveillance systems for monitoring emerging infectious diseases [J]. The lancet infectious diseases, 2014, 14 (2): 160－168.

2 JEREMY G, MATTHEW H M, RAJAN S P, et al. Detecting influenza epidemics using search engine query data [J]. Nature, 2009, 457 (7232): 1012－1014.

3 YUAN Q, NSOESIE O E, LV B, et al. Monitoring influenza epidemics in China with search query from Baidu [J]. Plos one, 2017, 8 (5): e64323.

4 DONALD R O, KEVIN J K, MARC P, et al. Reassessing Google flu trends data for detection of seasonal and pandemic influenza: a comparative epidemiological study at three geographic scales [J]. Plos computational biology, 2013, 9 (10): e1003256.

5 LUO Y, ZENG D, CAO Z, et al. Using multi－source web data for epidemic surveillance: a case study of the 2009 Influenza A (H1N1) pandemic in Beijing [C]. Toronto: 2010 IEEE International Conference on Service Operations and Logistics, and Informatics, 2010.

6 MAURICIO S, NSOESIE E O, MEKARU S R, et al. Using clinicians' search query data to monitor influenza epidemics [J]. Clinical infectious diseases, 2014, 59 (10): 1446－1450.

7 GLUSKIN R T, JOHANSSON M A, SANTILLANA M, et al. Evaluation of internet－based dengue query data: Google dengue trends [J]. Plos neglected tropical diseases, 2014, 8 (2): e2713.

8 ZHOU X, YE J, FENG Y. Tuberculosis surveillance by analyzing Google trends [J]. IEEE transactions on biomedical engineering, 2011, 58 (8): 2247－2254.

9 唐家博. 基于互联网数据的江苏手足口病的预警模型的研究 [D]. 南京: 东南大学, 2019.

10 BODNAR T, MARCEL S. Validating models for disease detection using twitter [C]. Brazil: The 22nd International Conference on World Wide Web, 2013.

11 LIU Q, LI Z, JI Y, et al. Forecasting the seasonality and trend of pulmonary tuberculosis in Jiangsu province of China using advanced statistical time－series analyses [EB/OL]. [2024－01－16]. https://www.tandfonline.com/doi/

full/10. 2147/IDR. S207809.

12 FLÁVIO F N, MONTEIRO A B S, TELLES P R, et al. Dynamic linear model and SARIMA: a comparison of their forecasting performance in epidemiology [J]. Statistics in medicine, 2001, 20 (20): 3051 - 3069.

13 TAN P N, STEINBACH M, KUMAR V. Introduction to data mining [M]. Chennai: Pearson Education India, 2016.

14 AMIN S, UDDIN M I, ALSAEED D H, et al. Early detection of seasonal outbreaks from twitter data using machine learning approaches [J]. Complexity, 2021 (3): 1 - 12.

15 张金宇. 基于气候和媒介及百度指数的多时间尺度登革热预测研究 [D]. 广州: 广东药科大学, 2020.

16 ZHOU Y, LI T, SHI J, et al. A CEEMDAN and XGBOOST - based approach to forecast crude oil prices [EB/OL]. [2024 - 01 - 16]. https: //www. hindawi. com/journals/complexity/2019/4392785/.

17 MENG D, XU J, ZHAO J. Analysis and prediction of hand, foot and mouth disease incidence in China using Random Forest and XGBoost [J]. Plos one, 2021, 16 (12): e0261629.

18 ARAMAKI E, MASKAWA S, MORITA M. Twitter catches the flu: detecting influenza epidemics using Twitter [C]. Edinburgh: The 2011 Conference on Empirical Methods in Natural Language Processing, 2011.

19 黄泽颖. 基于百度指数的传染病预测精准性探索——以广东省 H7N9 亚型禽流感为例 [J]. 中国人兽共患病学报, 2020, 36 (11): 962 - 968.

20 毛健, 赵红东, 姚婧婧. 人工神经网络的发展及应用 [J]. 电子设计工程, 2011, 19 (24): 62 - 65.

21 黄丽. BP 神经网络算法改进及应用研究 [D]. 重庆: 重庆师范大学, 2008.

22 王若佳. 融合百度指数的流感预测机理与实证研究 [J]. 情报学报, 2018, 37 (2): 206 - 219.

23 魏麟, 朱素玲, 胡晓斌. 基于 CEEMD - GRNN 组合模型的 HIV 感染病例数预测 [J]. 现代预防医学, 2022, 49 (6): 969 - 974.

24 杨德志. 广义回归神经网络在乙肝发病数时间序列预测中的应用 [J]. 计算机应用与软件, 2013, 30 (4): 217 - 219.

25 WANG G, WEI W, JIANG J, et al. Application of a long short - term memory neural network: a burgeoning method of deep learning in forecasting HIV incidence in Guangxi, China [EB/OL]. [2024 - 01 - 16]. https: //europepmc. org/article/pmc/6518582.

26 黄鹏. 基于机器学习的乙类传染病预测模型研究与实现 [D]. 成都: 电子科技大学, 2019.

27 PARWEZ M A, ABULAISH M, JAHIRUDDIN J. A social media time - series data analytics approach for digital epidemiology [C]. Melbourne: 2020 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI - IAT), 2020.

28 王若佳, 李培. 基于互联网搜索数据的流感监测模型比较与优化 [J]. 图书情报工作, 2016, 60 (18): 122 - 132.

29 SU K, XU L, LI G, et al. Forecasting influenza activity using self - adaptive AI model and multi - source data in Chongqing, China [EB/OL]. [2024 - 01 - 16]. https: //www. thelancet. com/article/S2352 - 3964 (19) 30546 - 8/fulltext.

30 赖晓蓥, 钱俊. ARIMA - LSTM - XGBoost 加权组合模型在肺结核发病趋势预测的研究 [J]. 现代预防医学, 2021, 48 (1): 5 - 9.

31 ANWAR M, KHOURY D, ALDRIDGE A P, et al. Using Twitter to surveil the opioid epidemic in North Carolina: an exploratory study [J]. JMIR public health and surveillance, 2020, 6 (2): e17574.

关于《医学信息学杂志》启用“科技期刊学术不端文献检测系统”的启事

为了提高编辑部对于学术不端文献的辨别能力，端正学风，维护作者权益，《医学信息学杂志》已正式启用“科技期刊学术不端文献检测系统”，对来稿进行逐篇检查。该系统以《中国学术文献网络出版总库》为全文比对数据库，可检测抄袭与剽窃、伪造、篡改、不当署名、一稿多投等学术不端文献。如查出作者所投稿件存在上述学术不端行为，本刊将立即做退稿处理并予以警告。希望广大作者在论文撰写中保持严谨、谨慎、端正的态度，自觉抵制任何有损学术声誉的行为。

《医学信息学杂志》编辑部