• 医学信息研究 •

基于信息流行病学的社会化媒体数据研究进展*

胡广宇

（中国医学科学院医学信息研究所/卫生政策与管理研究中心 北京 100020）

〔摘要〕 从信息流行病学角度阐述国内外基于社会化媒体数据开展的健康相关信息研究，着重介绍患者体验信息分析利用，从方法学和实践层面分析研究面临的问题与挑战，提出未来研究方向。

〔关键词〕 信息流行病学；社会化媒体；公共卫生监测；药物警戒；患者体验

〔中图分类号〕R－056 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2018.12.010

Progress in the Study on Socialized Media Data Based on Infodemiology *HU Guangyu, Institute of Medical Information/Center for Health Policy and Management, Chinese Academy of Medical Sciences, Beijing 100020, China*

〔**Abstract**〕 From the perspective of infodemiology, the paper expounds upon the implemented health－related information study based on socialized media data in and outside China. It focuses on the analysis and usage of information on patient experience, analyzes and investigates into confronting problems and challenges from the aspects of methodological and practice, as well as sets forward future direction of study.

〔**Keywords**〕 infodemiology; social media; public health surveillance; pharmacovigilance; patient experience

1 引言

近年来社会化媒体和社交网络服务在全球迅速兴起和广泛流行，大大拓宽传统健康相关研究的数据来源和媒介途径。与此同时全球范围内基于互联网信息开展健康相关研究的探索逐年增多，逐渐催生出信息流行病学（Infodemiology）这一新的学科和系列前沿研究方法。利用社会化媒体和社交网络服务数据，探索开展信息研究，是当前信息流行病学的一个重要应用领域。

2 概念与定义

2.1 信息流行病学

信息流行病学的概念由加拿大学者 Gunther Eysenbach 于 2002 年首次提出，将其定义为有关健康信息和虚假信息决定因素和分布研究的新兴学科及方法[1]。2009 年 Eysenbach 又将这一概念修订为研究电子媒介、互联网或人群中信息的分布及其决定因素，最终影响公共卫生和公共政策的科学[2]。传统的医学和健康相关研究方法如现场调查、队列研

〔收稿日期〕 2018－11－06

〔作者简介〕 胡广宇，博士，助理研究员，发表论文 20 余篇。

〔基金项目〕 中央级公益性科研院所基本科研业务费“医疗服务质量评价共性关键技术研究”（项目编号：2018PT33009）。

究、疾病登记等通常需要花费数年乃至数 10 年的时间才能产生具有政策影响的结果，时效性存在短板。信息流行病学的研究通过近乎实时的数据采集和分析大大拓展传统研究的内容和应用范围。基于供给侧和需求侧信息的研究方法与应用是信息流行病学研究的两大领域。前者的研究对象包括来自公开网站、博客、微博、社交网络服务等途径的公开发布信息，后者的研究对象包括来自互联网搜索引擎工具以及页面内容浏览监测等途径的用户使用行为特征数据[2]。

2.2 社会化媒体

2012 年美国国立医学图书馆将"Social Media"正式纳入医学主题词表（MeSH），中文医学主题词表（CMeSH）将其翻译为“社会化媒体”，定义为具有通过互联网创建和发布信息的能力及工具的平台。通常有 3 个特征，即内容由用户生成、创作者与阅读者之间高度互动、容易与其他网站进行整合。社会化媒体在中国更为常用的另一个表述是“社交媒体”。社交网络（Social Networking）是与社会化媒体既有关联又有区别的另一个概念，CMeSH 对其定义为：通过家庭、工作和其他利益相连的个体，也包括计算机通信促进的关系。社交网络在中国也是社交网络服务（Social Networking Service）的简称。社会化媒体与社交网络服务尽管在概念上相对较为明确，然而在具体服务平台的实际发展过程中随着用户需求和产品形态的快速演进和迭代两者往往并不完全以孤立的形式分别存在。从目前社会化媒体平台存在的产品形态来看可分为功能性细分平台和移动兴趣社区两类，社交网络平台可被视为功能性细分平台的一个子集，基于社交网络建立的社会化媒体平台兼具媒体与社交网络属性。在社会化媒体服务平台日新月异的大背景下针对社会化媒体的概念定义也在不断演进。

3 基于社会化媒体数据的健康相关信息研究

3.1 概述

2009 年谷歌公司和美国疾控中心合作在《自然》杂志发表利用公众搜索请求数据预测流感爆发的研究[3]，是信息流行病学研究的经典案例。此类研究属于需求侧信息的开发利用，用户的搜索请求数据属于内部数据，通常并不对外公开。而利用供给侧信息，尤其是基于社会化媒体数据开展研究，得益于公开领域信息的易得可用，受到更为广泛的关注。

3.2 疾病监测

疾病监测和公共卫生应对措施的制定与实施是信息流行病学的主要关注领域。针对 1999－2011 年期间利用社会化媒体数据开展疾病监测与预测研究的系统综述结果显示[4]，以 Twitter 为主要数据来源的研究居多，流感样病例（Influenza－like Illness，ILI）和 H1N1 是研究较多的疾病，研究普遍证实 Twitter 中有关 ILI 内容与疾控部门监测数据的相关性。此类研究表明对于采集及时可靠的疫情数据而言，实时的社会化媒体平台数据是值得关注的研究数据来源。2015 年 Fung 等人报道中国社会化媒体中用户对于法定传染病相关信息互动反应的分布分析结果[5]，然而基于中国数据开展疾病监测的研究报道仍不多见。此外针对 HIV[6]、埃博拉[7]以及中东呼吸综合征[8]等传染性疾病以及非自杀性自我伤害行为[9]的社会化媒体相关数据监测分析研究也是近年国际研究关注的焦点。

3.3 药物警戒

创新药物不良反应监测模式，解决药物相关问题，是信息流行病学在药物警戒领域的应用场景[10]。2014 年 Freifeld 等人的研究[11]证实 Twitter 上的药物不良反应事件与美国药监局不良事件报告系统中相关数据存在显著相关（$r = 0.75$，$p < 0.000\ 1$）。2017 年 Cocos 等人进一步利用递归神经网络（Recurrent Neural Network，RNN）模型开发出可扩展的深度学习方法，应用于 Twitter 数据的药物不良反应监测，取得了更好的监测和识别效果[12]。此外信息流行学研究还被扩展至药物开发和非法药物使用监测领域。2016 年美国梅奥诊所的研究人员报道利用患者对药物的在线评论数据开展药物再利

用分析的探索性研究结果[13]，揭示了社会化媒体数据对于药物开发的潜在价值；2017 年 Kazemi 等人针对近年来基于在线数据的非法药物使用监测相关研究的回顾性分析表明[14]，充分挖掘在线健康相关数据可为加强监管、推动解决非法药物使用这一全球性重大公共卫生问题提供新的思路。

3.4 干预评价

基于社会化媒体数据针对人群健康干预评价的研究更多关注特定人群对特定疾病或干预措施的评价和反馈，作为制定公共卫生干预措施的辅助性参考。Turner – McGrievy 和 Beets 在 2015 年发表社会化媒体中有关减肥话题的年度时间变化趋势分析结果[15]，发现假期中和假期后比假期前、冬季比夏季用户对有关减肥话题的关注度更高，认为通过“标签”在高关注度时段吸引用户关注减肥计划的话题可促进更为广泛减肥干预措施的实现。2017 年 Metwally 等人发表有关社会化媒体用户对癌症筛查态度的分析结果[16]，通过采集有关结肠镜检查、乳腺 X 光检查、巴氏涂片的社会化媒体话题数据发现用户对结肠镜检查的负面评论更多，对乳腺 X 光检查的正面评论多，而对巴氏涂片评论的情感偏好差异则不显著，进一步分析不同情感偏好内容的差异及及其用户分布和传播的特点，为癌症筛查更好地了解干预对象，设计适用性更好的干预策略提供独特视角。

4 基于社会化媒体数据的患者体验信息研究

4.1 概述

2013 年 Ronen Rozenblum 和 David W Bates 将以患者为中心的服务与互联网和社会化媒体的结合描述为“惊涛骇浪”[17]，认为将对卫生服务的供需双方产生深远影响。既往患者对供方的选择和评价主要依赖于声誉或朋友推荐[18]，对供方的服务质量、安全、服务体验也并无太多知晓途径。社会化媒体的出现，不仅为普通人群和患者提供公开表达意见和感受的平台，而且通过医疗服务评价等功能性平台以及患者或特定疾病的兴趣社区为普通公众更为深入和全面了解服务提供方提供重要途径。Greaves 等人针对此类源自社会化媒体的患者体验信息提出一个新的概念"cloud of patient experience"[19]，认为此类信息为患者体验和医疗服务质量研究提供全新视角。

4.2 方法学层面

Gonzalez – Hernandez 等人的研究认为[20]一般性社交网络平台（Generic Social Networks，GSNs）如 Facebook、Twitter 等，以及特定领域的网络平台如各种在线健康社区（Online Health Communities，OHCs），是研究数据的主要来源；得益于庞大的用户基数和海量服务记录数据，针对社会化媒体的数据挖掘研究近年来快速增长，而基于自然语言处理的有效噪音过滤机制和标准化研究概念映射是研究方法的技术关键。此外在将数据用于知识发现和服务监管之前仍有大量基础性研究工作有待完善。Greaves 等人[21]的研究证实情感分析技术对患者体验相关自由文本的处理可在患者有关医疗服务在线反馈的基础上给出合理准确的评估，且与传统患者体验评价调查结果具有较好相关性。Rastegar – Mojarad 等人探索建立患者体验专题语料库（Corpus of Patient Experience，COPE）[22]，此类研究为在英语国家利用社会化媒体数据深入开展患者体验研究提供重要的研究基础。

4.3 实践层面

2012 年以来有关于消费者在线医疗服务评价信息的研究在美国、英国、德国均有报道[23-25]。Greaves 等人对英国国家卫生服务体系官方在线医疗服务评价网站"NHS Choices" 中患者体验评价数据的研究[25]，以及 Bardach 等人利用美国商业性社交网络服务点评网站"Yelp" 涉及医院点评数据的研究[26]，均证实在医疗机构的层面消费者在线评价结果与临床诊疗效果和患者体验调查结果间存在广泛关联。2013 年 Timian 等人针对 Facebook 上美国纽约地区 40 家医疗机构账号点赞数的研究发现[27]，该指标与医疗机构住院 30 天心脏病死亡率负相关（$\beta = -92.88$，$P = 0.01$），与患者推荐意愿正相关

($\beta = 5.08$, $P = 0.04$)。但也有研究认为[28]，在线患者评价信息与患者体验评价结果以及质量安全指标关联性较弱，现有此类信息主要为患者就医选择提供参考，而对反映医疗质量作用有限。2016 年 Hao 等人首次将非监督机器学习技术中的文档主题生成模型隐含狄利克雷分布（Latent Dirichlet Allocation，LDA）应用于中国患者在线就医体验交流社区的医生点评信息分析[29]，发现中国患者就医体验反馈的热门话题分别涉及寻找医生的经历、医生的技术和态度、患者的感谢以及对于症状的描述等主题。相比美国同类研究结果而言，中美两国患者评价医生的表述内容存在文化差异。2018 年 Zhang 等人采用内容分析法研究中国患者有关就医反馈的负面在线评价内容[30]，结果显示妇产科和内科医生收到的负面评价居多，而患者负面评价中抱怨较多的内容包括诊疗时间不足、医生不耐烦、疗效不佳等，此外还发现不同特征人群对所接受服务的容忍度也有所差异。总体而言目前利用中国本土运营的社会化媒体信息开展的研究尚不多见。

5 问题与挑战

5.1 数据标准化与信息偏倚

当前全球范围内针对社会化媒体中健康相关信息处理的标准化和规范化研究仍较为薄弱[20]。医学和健康的相关概念本身就具有专业性特征，而在使用社会化媒体的过程中用户在不同平台对相关概念的不同表达和表述形式，包括对表情、动态图片等非文本化语言的使用，为数据的标准化处理提出更为复杂和更多待解决的问题。与此相关，对于健康领域文本信息的噪声过滤、准确分类、语义解释等问题，现有技术方法的解决效率和准确性、可靠性仍有待进一步提升，尤其是针对中文领域的此类基础性研究，目前仍缺乏重要进展。信息流行病学研究中同样存在偏倚，且由于数据量倍增，偏倚效应也可能被越加放大。基于社会化媒体数据的研究中最为普遍存在的是选择偏倚，一方面来自社会化媒体平台本身，不同类型服务平台的用户特征各有不同，不同来源数据对一般人群的代表性需要具体评估；另一方面来自研究过程中对目标研究内容的信息和数据抓取采集有偏或倚不够全面，导致信息中的报告偏倚和调查者偏倚。

5.2 隐私保护与伦理学挑战

尽管各国政府都有提升公共卫生和卫生服务监测效率的迫切需求，但将来自互联网和社会化媒体数据的监测手段和方法整合到传统的官方监测体系中仍面临可靠的自动化分析技术缺乏、政府相关部门接受度差异以及缺乏共识性看法等问题[31]。此外更具有争议且无法回避的是涉及信息隐私保护和用户知情同意的问题。2017 年 Sinnenberg 等人针对以 Twitter 为数据源的健康相关研究回顾性分析发现[32]，部分研究认为由于用户数据的公开性质，以及用户对于平台服务条款的事前同意，声明免除研究开展的伦理审查和知情同意。尽管全球几乎所有的社会化媒体平台在用户协议中都有关于用户隐私保护和数据利用的相关说明条款，但对于第 3 方包括研究用途的数据使用规则，迄今并无范式也缺乏明确的解释。Vayena 等人认为此类研究产生新的前所未有的伦理学挑战[33]，当前需要制定通用性伦理指南指导此类研究的开展[32]。

6 结论

信息流行病学是当前交叉学科研究的前沿，新的应用场景开发，信息的有效聚合，特征数据的提取分析利用，是基于网络和社会化媒体数据开展此类研究，有待深入探索的领域。与计算机科学、心理和行为科学等领域的合作，以及对其他学科成熟研究方法的应用，是有待进一步探索的方向。建立针对基于社会化媒体数据研究的标准化报告规范是未来有待推进的一项关键性工作，目前针对社会化媒体数据的研究，不同的文献报告在数据来源说明、数据采集方式、模型分析方法、数据质量评价、伦理声明等方面不尽相同。标准化报告规范的建立和使用将有助于推动研究的规范化和结果的重复可比。就面向政府和公众提供疾病监测服务而言，尽管当前针对基于互联网信息开展疾病监测的

价值几何仍缺乏可靠的系统性评估研究证据[31]，但对非洲、亚洲、南美等部分传染性疾病负担较重的发展中国家，由于实施传统监测的工作基础和条件较为缺乏，考虑先行利用来自互联网和社会化媒体的相关信息作为开展传统监测的补充，对于预防疾病爆发，降低公共卫生风险是具有现实意义的选择。即使对中国已建立法定传染病报告制度和成熟疾病监测网络的国家系统，从信息流行病学的视角探索新的疾病监测途径和方法，未来也仍具有广阔应用前景。此外从信息流行病学视角，利用社会化媒体相关信息，探索开展第 3 方患者体验评价，是未来可行性较好且极有价值的一个研究方向。此类研究不仅可为公众就医选择提供更有价值和更为可靠的参考信息，同时可为政府部门推进医疗服务监管提供新的途径与方法。尤其是与传统患者调查研究的结合，将为国内现有的医疗服务质量评价体系提供更为全面和丰富的视角与信息，有利于推进以患者为中心的理念，得到更为广泛的理解认同与实践实现。

参考文献

1 Eysenbach G. Infodemiology: the epidemiology of (mis) information [J]. The American Journal of Medicine, 2002, 113 (9): 763 - 765.

2 Eysenbach G. Infodemiology and Infoveillance: framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the internet [J]. Journal of Medical Internet Research, 2009, 11 (1): e11.

3 Ginsberg J, Mohebbi M H, Patel R S, et al. Detecting Influenza Epidemics Using Search Engine Query Data [J]. Nature, 2009, 457 (7232): 1012 - 1014.

4 Guy S, Ratzki - Leewing A, Bahati R, et al. Social Media: a systematic review to understand the evidence and application in infodemiology [M]. Heidelberg: Springer, 2012: 1 - 8.

5 Fung I C, Hao Y, Cai J, et al. Chinese Social Media Reaction to Information about 42 Notifiable Infectious Diseases [J]. Plos One, 2015, 10 (5): e126092.

6 Young S D, Rivers C, Lewis B. Methods of Using Real - time Social Media Technologies for Detection and Remote Monitoring of HIV Outcomes [J]. Preventive Medicine, 2014 (63): 112 - 115.

7 Odlum M, Yoon S. What Can We Learn About the Ebola Outbreak from Tweets? [J]. American Journal of Infection Control, 2015, 43 (6): 563 - 571.

8 Song J, Song T M, Seo D C, et al. Social Big Data Analysis of Information Spread and Perceived Infection Risk During the 2015 Middle East Respiratory Syndrome Outbreak in South Korea [J]. Cyberpsychol Behav Soc Netw, 2017, 20 (1): 22 - 29.

9 Moreno M A, Ton A, Selkie E, et al. Secret Society 123: understanding the language of self - harm on instagram [J]. Journal of Adolescent Health, 2016, 58 (1): 78 - 84.

10 Sarker A, Ginn R, Nikfarjam A, et al. Utilizing Social Media Data for Pharmacovigilance: a review [J]. Journal of Biomedical Informatics, 2015 (54): 202 - 212.

11 Freifeld C C, Brownstein J S, Menone C M, et al. Digital Drug Safety Surveillance: monitoring pharmaceutical products in twitter [J]. Drug Saf, 2014, 37 (5): 343 - 350.

12 Cocos A, Fiks A G, Masino A J. Deep Learning for Pharmacovigilance: recurrent neural network architectures for labeling adverse drug reactions in Twitter posts [J]. Journal of the American Medical Informatics Association, 2017, 24 (4): 813 - 821.

13 Rastegar - Mojarad M, Liu H, Nambisan P. Using Social Media Data to Identify Potential Candidates for Drug Repurposing: a feasibility study [J]. JMIR Research Protocols, 2016, 5 (2): e121.

14 Kazemi D M, Borsari B, Levine M J, et al. Systematic Review of Surveillance by Social Media Platforms for Illicit Drug Use [J]. Journal of Public Health, 2017, 39 (4): 763 - 776.

15 Turner - Mcgrievy G M, Beets M W. Tweet for Health: using an online social network to examine temporal trends in weight loss - related posts [J]. Transl Behav Med, 2015, 5 (2): 160 - 166.

16 Metwally O, Blumberg S, Ladabaum U, et al. Using Social Media to Characterize Public Sentiment Toward Medical Interventions Commonly Used for Cancer Screening: an observational study [J]. Journal of Medical Internet Research, 2017, 19 (6): e200.

17 Rozenblum R, Bates D W. Patient - centred Healthcare, Social Media and the Internet: the perfect storm? [J]. BMJ Quality & Safety, 2013, 22 (3): 183 - 186.

18 Bates D W, Gawande A A. The Impact of the Internet on Quality Measurement [J]. Health Aff (Millwood), 2000, 19 (6): 104 – 114.

19 Greaves F, Ramirez – Cano D, Millett C, et al. Harnessing the Cloud of Patient Experience: using social media to detect poor quality healthcare [J]. BMJ Quality & Safety, 2013, 22 (3): 251 – 255.

20 Gonzalez – Hernandez G, Sarker A, Connor K O, et al. Capturing the Patient's Perspective: a review of advances in natural language processing of health – related text [J]. Yearbook of Medical Informatics, 2017, 26 (1): 214 – 227.

21 Greaves F, Ramirez – Cano D, Millett C, et al. Use of Sentiment Analysis for Capturing Patient Experience From Free – Text Comments Posted Online [J]. Journal of Medical Internet Research, 2013, 15 (11): e239.

22 Rastegar – Mojarad M, Ye Z, Wall D, et al. Collecting and Analyzing Patient Experiences of Health Care From Social Media [J]. JMIR Res Protoc, 2015, 4 (3): e78.

23 Mclennan S, Strech D, Reimann S. Developments in the Frequency of Ratings and Evaluation Tendencies: a review of german physician rating websites [J]. Journal of Medical Internet Research, 2017, 19 (8): e299.

24 Gao G G, Mccullough J S, Agarwal R, et al. A Changing Landscape of Physician Quality Reporting: analysis of patients' online ratings of their physicians over a 5 – Year period [J]. Journal of Medical Internet Research, 2012, 14 (1): e38.

25 Greaves F. Associations Between Web – Based Patient Ratings and Objective Measures of Hospital Quality [J]. Archives of Internal Medicine, 2012, 172 (5): 435.

26 Bardach N S, Asteria – Penaloza R, Boscardin W J, et al. The Relationship Between Commercial Website Ratings and Traditional Hospital Performance Measures in the USA [J]. BMJ Quality & Safety, 2013, 22 (3): 194 – 202.

27 Timian A, Rupcic S, Kachnowski S, et al. Do Patients "Like" Good Care? Measuring Hospital Quality via Facebook [J]. American Journal of Medical Quality, 2013, 28 (5): 374 – 382.

28 Emmert M, Meszmer N, Schlesinger M. A Cross – sectional Study Assessing the Association Between Online Ratings and Clinical Quality of Care Measures for US Hospitals: results from an observational study [J]. BMC Health Services Research, 2018, 18 (1): 82.

29 Hao H, Zhang K. The Voice of Chinese Health Consumers: a text mining approach to web – based physician reviews [J]. Journal of Medical Internet Research, 2016, 18 (5): e108 – e120.

30 Zhang W, Deng Z, Hong Z, et al. Unhappy Patients Are Not Alike: content analysis of the negative comments from china's good doctor website [J]. Journal of Medical Internet Research, 2018, 20 (1): e35.

31 Velasco E, Agheneza T, Denecke K, et al. Social Media and Internet – Based Data in Global Systems for Public Health Surveillance: a systematic review [J]. Milbank Q, 2014, 92 (1): 7 – 33.

32 Sinnenberg L, Buttenheim A M, Padrez K, et al. Twitter as a Tool for Health Research: a systematic review [J]. American Journal of Public Health, 2017, 107 (1): e1 – e8.

33 Vayena E, Salathe M, Madoff L C, et al. Ethical Challenges of Big Data in Public Health [J]. PLoS Comput Biol, 2015, 11 (2): e1003904.