中国科学: 信息科学 2020 年 第 50 卷 第 12 期: 1882–1902
SCIENTIA SINICA Informationis
论文

基于大规模结构化病例数据的新型冠状病毒传播特征和感染人群分析

黄振华[1,2*], 王振宇[1*], 江莉[3], 张睿[1], 雷昶[1], 刘星炜[2], 谢晓辉[2]

1. 华南理工大学软件学院, 广州 510006, 中国
2. Department of Information and Computer Science, University of California Irvine, Irvine 92617, USA
3. 安徽医科大学第一附属医院药剂科, 合肥 230032, 中国
* 通信作者. E-mail: sezhhuang@mail.scut.edu.cn, wangzy@scut.edu.cn

收稿日期: 2020–02–17; 修回日期: 2020–03–01; 接受日期: 2020–03–20; 网络出版日期: 2020–05–07

国家自然科学基金面上项目 (批准号: 61876207)、广东省重点领域研发计划项目 (批准号: 2019B010154004)、广东省基础与应用基础研究基金项目 (批准号: 2019A1515011792)、广州市产业技术重大攻关计划项目 (批准号: 201802010025) 和广州市高校创新创业平台建设项目重点项目 (批准号: 2019PT103) 资助

摘要 2020 年年初, 新型冠状病毒感染的肺炎 (COVID-19) 爆发, 中国采取了全面严格的防控举措全力抗击疫情. 地方疫情指挥部门及时通报疫情感染数据, 有助公众了解疫情的发展, 及时做好防护措施. 各地患者病例详情数据主要以文本形式记录, 信息描述复杂, 且各省市汇报的格式各异, 处理难度较大. 我们面向全国湖北省外近二分之一匿名的患者病例详情数据, 提出了应用自然语言处理技术, 辅助病例数据结构化的方法. 该方法可以在标记样本较少的情况下, 借助预训练模型, 准确有效地提取出病例文本中的关键信息. 通过对较大规模患者结构化病例数据的挖掘, 本文详细分析了新型冠状肺炎总体发病性别和年龄分布特点、主要感染原因、潜伏期特点及疫情趋势等特征. 由于潜伏期等时间延迟的存在, 确诊人数往往不能反映一个地区的真实感染情况, 结合出行大数据, 本文提出了一个合理推断武汉市等城市实际感染人数的方法. 该方法有助于人们提前估计地区疫情发展情况, 及早采取防护措施. 也可以辅助地方相关部门科学决策, 尽早调度医务人员和分配医疗资源.

关键词 新型冠状病毒, 结构化病例, 自然语言处理, 预训练模型, COVID-19 传播特征, 出行大数据

1 简介

2020 年年初, 爆发新型冠状病毒感染的肺炎 (COVID-19), 引起了全国乃至全世界的关注. 新型冠状病毒给疫区患者的生命健康带来不可估量的损失, 关于新型冠状病毒的起源也尚在探究中, 然而疫情的爆发是一个渐进的过程, 在初期就显示许多特征, 如果我们在疫情早期利用大数据分析技术方法, 就能较早地评估疫情的严重性, 把握阻击疫情的最佳时期, 尽早地准备和防范.

引用格式: 黄振华, 王振宇, 江莉, 等. 基于大规模结构化病例数据的新型冠状病毒传播特征和感染人群分析. 中国科学: 信息科学, 2020, 50: 1882–1902, doi: 10.1360/SSI-2020-0029
Huang Z H, Wang Z Y, Jiang L, et al. Analysis of COVID-19 spread characteristics and infection numbers based on large-scale structured case data (in Chinese). Sci Sin Inform, 2020, 50: 1882–1902, doi: 10.1360/SSI-2020-0029

2 月份前, 湖北省武汉市的疫情最为严重复杂, 由于医疗物资、试剂缺乏、检测能力受限等因素影响, 积累了大量感染却未能及时确诊的病例, 导致后期确诊病例爆发式增长. 当时确诊病例等数据并不能反映湖北省武汉市真实疫情情况, 武汉市有约 500 万人流出[1)], 其中感染人员也随之涌向全国各地, 流入到外省的患者数据和真实详情更容易得到, 该数据从侧面反映了湖北省武汉市的疫情. 本文汇总了湖北省外的 11 个省 (直辖市、自治区) 2 月 10 日以前的病例详情数据, 对 4815 名患者的匿名病例详情数据做分析, 约占 2 月 10 日前除湖北外各省确诊患者的二分之一. 为了保证分析的可靠性, 我们尽量确保省市数据的完整性, 其中包括安徽省、天津市、山东省、陕西省、广西壮族自治区、海南省的完整病例数据和河南省、湖南省、广东省、江苏省部分城市的完整数据. 对于非结构化的病例详情文本数据, 我们提出一种自然语言处理方法, 辅助病例文本数据的结构化, 辅助提取病例的结构化信息, 如性别、年龄、地区、发病原因、发病日期、感染日期、就诊日期、确诊日期等信息. 相比之前的结构化方法 [1,2], 该方法借助预训练模型, 在标注样本较少的数据下也能有较好的表现, 大量节省了分析数据的人力、物力和时间, 可以应用于大规模病例详情数据分析[2)]. 相比于国内外一些团队的工作, 更加智能化.

病毒感染机制非常复杂, 受到人群体质、接触时间、接触方式等的影响. 武汉市疫情初期情况复杂, 积累了大量感染病例, 表面上看到的确诊数据和真实的感染情况相差较大, 给人造成了疫情并不严重的错觉. 因此, 本文通过结构化数据和出行大数据估计封城前武汉市的感染人数和人群感染率. 相比于帝国理工大学 (Imperial College London) [3]、美国东北大学 (Northeastern University) [4] 等实验团队提出的模型, 更具可信度和合理性, 大幅度降低了预测误差. 实际数据结果证明, 较小的人群感染率, 在不加控制的情况下, 在复杂社会网络上, 也能造成大面积的感染.

至本文工作完成时 (2 月 10 日), 据我们了解, 本文所使用的数据是最大规模的新型冠状病毒结构化病例详情文本数据 (非结构化的病例数据分析不纳入对比), 是首项采用省市完整的病例详情数据分析新型冠状病毒传染特点的工作. 之前公布的研究大多基于全国各省市零碎的数据, 未能基于一个地区的完整数据去分析地区传染特点. 通过其他城市的完整的病例数据, 才能够较为合理地估计武汉市等城市的真实感染人数. 总结本文贡献如下:

(1) 提出了人工智能辅助的病例结构化方法, 应用深度学习模型和预训练方法, 在标记样本较少的情况下, 也能得到比较精确的结果, 辅助相关人员在较短时间内完成大规模病例详情数据的分析.

(2) 基于新型冠状病毒结构化的病例文本数据, 详细地分析此次病毒传染人群的性别年龄特点、感染原因等特征, 求解符合新型冠状病毒的潜伏期、就诊延迟、确诊延迟的数学分布以及基础再生数 $R0$, 通过数据挖掘疫情爆发的部分原因.

(3) 结合出行大数据和结构化数据, 提出了在疫情早期, 合理估计一个城市实际感染人数的方法, 相比现有的工作, 大幅度降低估计误差, 更具参考价值.

2 相关工作

COVID-19 的爆发, 引起了国内外各界人士的关注和支持. 研究人员从不同角度研究新型冠状病毒, 为病毒特征理解、临床治疗、疫苗研发、疫情防控提供参考 [5~7][3)]. 中国疾病预防控制中心、华中科技大学等机构联合发表在 *Nature* 上的文章 [6] 揭示新型冠状病毒早期患者的病例临床详情, 对该

1) https://www.sohu.com/a/369170126_120207617.

2) 本文所指大规模是相对于传统医学个案病例分析而言的, 并非指信息科学中的并行计算、分布式计算等大数据技术.

3) https://new.qq.com/omn/20200212/20200212A04QFW00.

患者发病症状 (发烧、头晕和咳嗽) 和胸部 X 线 (双侧弥漫性斑块状的模糊阴影) 进行分析, 并推演了 COVID-19 病毒的进化史 [3)].

在患者临床表现方面, 高福院士团队发表首篇论文, 根据 425 例早期患者, 估计 COVID-19 潜伏期平均 5.2 天, $R0$ 值估计为 2.2, 感染人数每 7.4 天翻一倍, 并分析了早期患者和华南海鲜市场的关系 [8]. *Lancet* 报道了武汉金银潭医院、中日友好医院等机构的研究工作, 描述了新型冠状病毒感染患者的临床和医学影像学特征, 以及治疗方法, 并比较了重症和轻症患者的临床特征, 研究结果通报世界各国 [9]. Holshue 等 [10] 在 *New England Journal of Medicine* (NEJM) 上报道了美国首例患者治愈的过程, 使用尚未上市、原用于治疗埃博拉冠状病毒的 Remdesivir (瑞德西韦) 治愈了重症患者. 武汉金银潭医院的张丽、上海交大瑞金医院的张欣欣等 [11] 在 *Lancet* 报道了 99 例武汉患者的临床表现, 指出通过影像学检查, 75% 的患者出现双侧肺炎, 14% 的患者有肺部斑驳和浑浊, 病毒可导致严重、致命的呼吸系统疾病, 如急性呼吸窘迫综合征 (ARDS). 国家相关部门及时汇总更新治疗方案, 根据患者症状和发病特征分级治疗, 发布了新型冠状病毒感染肺炎的诊疗方案[4)]. 钟南山院士最早公开肯定新型冠状病毒会人传人, 其团队 30 名研究人员对包括武汉市 1099 临床病例数据进行手动统计分析, 研究发现患者平均年龄 47 岁, 41.9% 是女性, 71.80% 的人与来自武汉的人有过接触, 在研究样本中, 死亡率为 1.36%, 潜伏期中位数为 4 天 [12]. 这种分析方式细致, 但需要大量人力物力, 而且有可能会出现一些人工产生的笔误, 而本文采用人工智能辅助结构化病例的方法, 能够大幅度提高数据处理和分析能力.

在预测和估计疫情爆发上, 英国帝国理工大学提出使用武汉市机场发客流量和境外确诊人数估计武汉市感染人数 [3], 这种简易模型认为境外确诊的病人数 = 武汉总感染人数 × 武汉人每天坐飞机出境的概率 × 可以供武汉感染者出境且在境外才被发现的时间窗口长度 (天数). 该团队认为武汉市共有 1900 万人口, 估计从感染到确诊时间窗口大约十天时间, 预测 1 月 18 日武汉市具有 4000 病例. 美国东北大学研究组的 Alessandro 教授等 [4] 将全球分为若干区域, 充分考虑了任意两个区域之间的人群流动性数据. 该团队同样根据境外已经确诊的人数 D、估计的感染人数 N, 应用贝叶斯 (Bayes) 定理: $P(N) \times P(D|N) = P(D) \times P(N|D)$, 估计 $P(N|D)$, 认为 2020 年 1 月 20 日武汉市感染人数为 4050 人. 这些方法具有一定参考性, 但只使用了机场人流, 患者样本规模较小, 且抽样人群具有偏向性, 难以表现整体特征. 我们通过出行大数据和结构化数据估计武汉市感染人数, 更具合理性.

3 患者病例数据结构化

3.1 病例详情数据示例

不同的省份、同一省的不同城市, 记录病例的方式、详尽程度各异, 部分省市的病例详情数据, 样例如表 1. 为结构化患者匿名病例详情数据, 我们需要从中提取出患者假名、性别、年龄、地区、感染时间、发病时间、就诊时间、确诊时间、是否旅居武汉、旅居武汉原因、是否密切接触患者等信息. 注意并非所有病例都存在这些完整的数据, 有些病例并没有记录明确的感染或发病时间等. 感染时间、发病时间、就诊时间、确诊时间的描述分别详见下面的定义 1~4.

定义1 (感染时间) 病例接触传染源的合理估计时间.

患者真实的感染时间往往只能通过其行程和活动时间来估计. 感染时间包括去武汉市、湖北省等疫区, 接触亲密患者等被传染的时间. 除非延迟时间过长, 或者患者发病前有其他接触感染源的行为.

4) http://www.gov.cn/zhengce/2020-02/05/content_5474852.htm.

表 1 部分省市匿名病例详情数据样例

Table 1 Samples of anonymous COVID-19 cases from different places

Province/City	Case examples
Anhui	xxx, male, 45 years old, Xiantao City, Hubei Province ···. When the patient returned to his hometown by car from Hubei on Jan. 10th, he first hung out for 3 hours on Wuhan Hanzheng Street, and then returned to xxx town of Mengcheng on the 11th. He began to cough, mainly dry cough, on the 24th, and transferred to the First People's Hospital of Mengcheng for treatment on the 26th.
Guangxi	xxx, female, 24 years old, xxx from Guilin, is the wife of patient xxx who was confirmed on Feb. 1st. On Jan. 27th, she and xxx returned to Guilin via Wuhan. On the 25th, she showed symptoms such as fever and sputum appeared. On Feb. 1st, she was hospitalized in the Third People's Hospital. On Feb. 4th, she was tested positive ····.
Shenzhen	36 years old male patient, resident in Shenzhen Nanshan. He drove to Ezhou, Hubei on January 20th and returned to Shenzhen on the 25th. He began to show symptoms on February 1st and was hospitalized on February 3rd. He is now in a stable condition ····.

例如: "病例 xxx, 12 月 25 日去过武汉, 返乡后 1 月 15 日和确诊患者接触", 取 1 月 15 日被感染. 如果患者在武汉市待了 3 天. 例如: "病例 xxx, 1 月 8 日 ∼14 日期间去过武汉出差". 最晚可能的感染时间为 1 月 14 日, 本文分析潜伏期时采用最晚可能感染时间, 因此潜伏期是略有低估的.

定义2 (发病时间) 病例最早出现症状的时间. 症状包括感到不适、干咳、咳痰、咽痛、发热、四肢无力、肌肉酸痛、腹泻等.

定义3 (就诊时间) 病例发病后到县级以上医院就诊时间.

本文所指就诊时间不同于初次看医生的时间, 而是指去县级以上医院就诊后病例被隔离观察, 基本上无法继续传染他人的时间. 隔离时间一般早于或等于就诊时间. 在村医、社区医院、诊所就诊不等于就诊时间, 因为这些小医院 (诊所) 不具有隔离手段, 患者还能继续传染其他人, 不但起不到治疗效果, 反而进一步扩散了疫情, 同时患者自身也耽误了治疗最佳时机.

定义4 (确诊时间) 病例被医院确诊或官方通报确诊的时间.

一般而言, 从就诊到确诊也需要一段时间, 最初平均需要 2∼3 天. 地方卫生健康委员会官方通报一般在医院确诊的第 2 天, 如果两者时间均存在, 取最早确诊时间, 即医院确诊时间.

3.2 数据结构化方法

对于输入病例文本数据, 首先将病例文本进行切分, 将文字序列转化为若干词或者字的符号 (token), 如图 1, 得到 token: "[CLS]"、"病"、"例"、", "、"34"、"岁" ·····, 其中, "[CLS]" 表示病例的起始符号.

对这些 token 进行表示学习, 分别对 "病", "例", ", ", "34", "岁" 用一个低维、连续、稠密向量表示, 每个向量表达了这些词语在欧式空间里的语义特征. 例如 "病" 和 "例" 在表示学习欧氏空间距离时会比较接近. 输入: $X \in \mathbb{R}^{N\times F}$, 其中 F 是 token 的表示学习向量维度, 本文选取 768 维, N 是一个句子的长度. 考虑到词语的位置会隐含部分信息, 我们增加了词语所在病例位置的表示特征.

Transformer (互感神经网络) [13]5) 接受自然语言描述的句子文本输入, 对句子的语义特征、位置特征等进行学习. 互感神经网络自注意力层使用自注意力机制学习 token 之间语言上的相互依赖关系, 每个 token 都能 "感受" 到其他 token 对它的影响和依赖关系. 将输入特征 X 映射成 Q, K 和 V, 并

5) 目前 Transformer 还没有统一的中文翻译, 由于每个 token 相互感知, 本文称之 "互感神经网络".

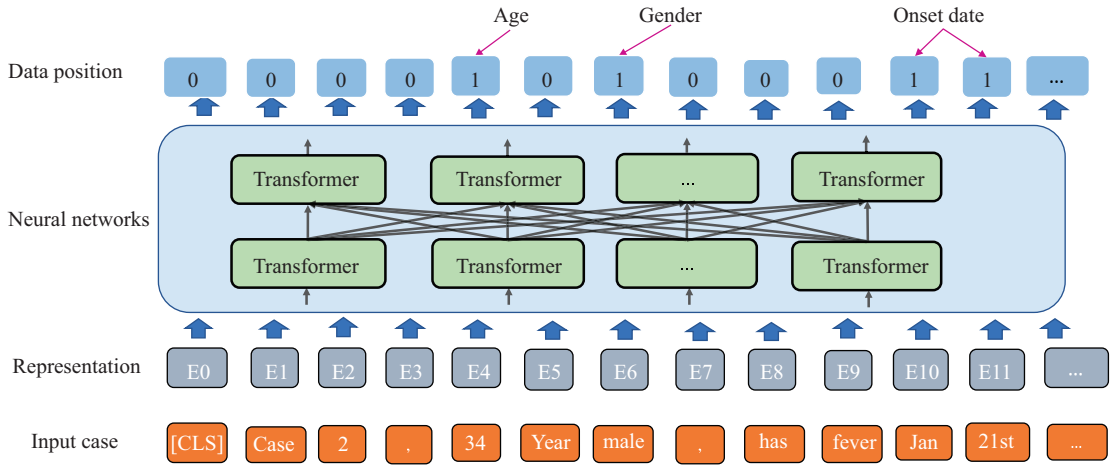


图 1 (网络版彩图) 深度学习模型提取病例患者性别、年龄、日期信息

Figure 1 (Color online) The deep learning model to extract patient information including: gender, age and date

学习 Q 与 K 之间的相互关系. 注意力机制采用公式 $\text{attn}^i = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$ 计算.

多头注意力机制把自注意力机制的特征 h 连接在一起: $\text{Concat}(\text{attn}^0, \text{attn}^1, \ldots, \text{attn}^h)W^{\text{mh}}$, 其中 W^{mh} 是带学习参数.

自注意力层的输出输入到前馈神经网络中, 并采用残差网络[14] 增强神经网络学习能力. $\text{FFN}(h) = \text{ReLu}(hW_1 + b_1)W_2 + b_2$, 其中, W_1 和 W_2 是神经网络参数, b_1 和 b_2 是偏置参数.

输出: $H \in \mathbb{R}^{N\times F}$. 每个词语 x_i 对应一个互感神经网络表示的向量 h_i.

最后, 对输出向量 h_i 加一个全连接层, 做一个分类任务, 得到模型输出的类别 $\hat{y}_i$, 损失函数交叉熵为 $L = -\sum^M \sum_i^C y_i \log \hat{y}_i$, 其中 M 是训练样本数量, C 是类别数量.

对于日期提取等相对复杂的任务, 由于每个目标 token 之间可能还会依赖前一时刻的 token, 例如, “25”、“日”, “日” 会考虑到前面的 “25”. 因此我们在输出向量后再使用一个 CRF 层, CRF 的目标函数不仅考虑 h_i 的特征, 还考虑了标签序列的转移特征, 在长目标序列提取上可能会更加有效, 若表现有提升, 则取增加 CRF 层后的结果.

3.3 模型训练

本文对年龄、性别、地区、日期等提取任务分开独立做识别, 联合做识别的方法留待后面的工作. 对于每个任务, 我们选择 1000 个样本, 其中 800 个做训练集, 200 个做测试集, 以一个极小的学习率 2E−5, 3E−5, 5E−5 训练 5~15 轮, 取最好的结果. 训练前, 使用文献 [15] 预训练的中文词向量初始化模型. 实验结果表明, 在提取病例年龄和性别测试集上准确率 (accuracy) 可以高达 100%. 对于性别的提取也可以采用文本分类的方式. 在地区的提取上准确率高达 99.5%. 在判断日期是否为感染时间、发病时间、就诊时间、确诊时间上准确率分别为 97%, 98%, 95%, 100%. 与 LSTM+CNN+CRF [16] 方法进行对比, LSTM+CNN+CRF 的准确率分别为 86%, 91.5%, 82%, 96% [6]. 基于循环神经网络 LSTM 的方法会存在较为严重的时间提取片段不完整或过多的问题, 例如 “2 月 1 日” 只提取了 “2 月”.

以上工作中, 确诊时间比较容易提取, 一般以 “x 月 x 日确诊” 或 “核酸检测阳性” 这样的形式给出, 准确率可以达到 100%. 模型通过识别患者是否去了疫区或是否亲密接触患者的描述来判断感染

6) 在 LSTM 的基础上增加预训练模型, 可以提高准确率.

时间, 例如 “x 月 x 日从武汉出差返回”、“x 月 x 日和确诊病例接触” 等. 因部分噪声日期的存在, 感染时间的提取和判断相对难一些. 但即便在预测错误的部分案例中, 预测感染时间和标记的感染时间也非常接近. 例如: 病例文本数据 “1 月 19 日患者到 xx 县参加婚礼, 与同行人住酒店, 同住人员有 1 人来自武汉. 1 月 20 日乘大巴车返回泗县, $\cdots$, 1 月 22 日由姐夫驾车回老家, $\cdots$, 1 月 26 日, $\cdots$, 1 月 30 日发病 $\cdots$”, 模型提取感染时间为 1 月 20 日, 而标记感染时间为 1 月 19 日. 如果具体感染时间不明确, 存在一个感染时间段, 则预测该时间段范围内任意一个时间皆为正确. 例如 1 月 16~17 日去武汉市旅游, 感染时间提取为 16 日或 17 日皆算正确. 对于发病时间, 会存在多次发病时间的问题, 患者可能存在 1 月 22 日咳嗽和 1 月 23 日发热, 其记录时间相隔一般不大, 只要模型完整输出其中一个即为正确. 就诊时间准确率相对较低, 因为一部分患者存在多次就诊和转诊, 以及去村医、社区医院就诊的情况, 这让模型相对较难判断. 如果文中出现多次就诊时间且间隔不超过 1 天, 完整提取其中一个则视为正确. 考虑到少部分患者轨迹和病例描述的复杂性, 对于准确率不足的时间提取, 由人工智能辅助给出答案, 最后交由专业者进行纠正和选择. 结构化的数据和人工记录皆会有极个别的错误, 但并不影响整体分布的分析. 在感染原因上, 采用文本分类进行学习, 在是否具有旅居武汉史, 是否亲密接触患者上准确率分别可以达到 97.5% 和 98%. 该方法类似一个生产线模式, 人工智能是生产线, 人工或专家是生产线上的质量检测员, 纠正人工智能的错误, 并且反馈给人工智能模型, 模型在收到反馈后不断地增强训练效果.

3.4 结构化病例结果

结构化后的病例数据, 如下: 是否旅居武汉, 包括长期旅居武汉 (长期居住武汉或者流动人口)、从武汉出差返回、去武汉出差、探亲、路过. 接触患者原因包括: 家人 (夫妻、父母、儿女、兄弟姐妹、亲戚等)、同事、聚餐、看望病人等, 所提结构化的模型, 并不对细化的接触患者原因 (家人、聚餐等) 做出区分, 细化后的接触患者原因样本数量有限且极不均衡, 很难训练出接近 100% 准确率的模型, 而是后期采用关键词提取的方法, 提取出描述接触患者信息的关键词, 以供专家分析使用.

此外, 我们发现, 是否跨区域就诊等特殊情况对疫情也有很重要的影响. 由于特殊情况的样本也较少, 难以训练高准确率的机器学习模型. 使用模式匹配方法, 提取旅行时间的情况, 如果旅行时间晚于发病时间, 说明该患者患病出行, 很可能异地就诊了, 再人工确认. 根据关键词匹配, 社区医院、诊所、村医等关键词判断是否去了非隔离措施的小医院就诊.

备注关键词包括: 是否去过诊所 (社区医院、村镇医院)、是否存在发热症状却返乡 (异地就医)、是否存在外地旅居史、是否存在聚餐 (聚会) 等. 得到结构化的患者病例详情数据如图 2.

4 结构化病例数据分析

4.1 感染原因分析

染病原因主要包括 “旅居武汉”、“接触患者” 和 “社区感染”. 还有一些因统计工作等原因未明确标明. 旅居武汉包括常住武汉、在武汉工作、去武汉探亲、旅游和出差等. “接触患者” 包括患者是家人、亲属、朋友、同事, 或在聚餐、宴会、看望病人等场合亲密接触患者. “社区感染” 指去超市购物、商场逛街、外地旅居、饭馆就餐等原因被感染.

湖北外的病例中, 从图 3 中可以看到, 总体上来看, 具有武汉旅居史的病例占据 62.6%, 是最主要感染原因, 武汉市输入病例也是各省市疫情爆发的主要原因. 但不同省市传染源二次传播的感染程度不同, 天津市、山东省、河南省接触患者二次传播占比也相对较高, 其他省份输入病例占据比重均超

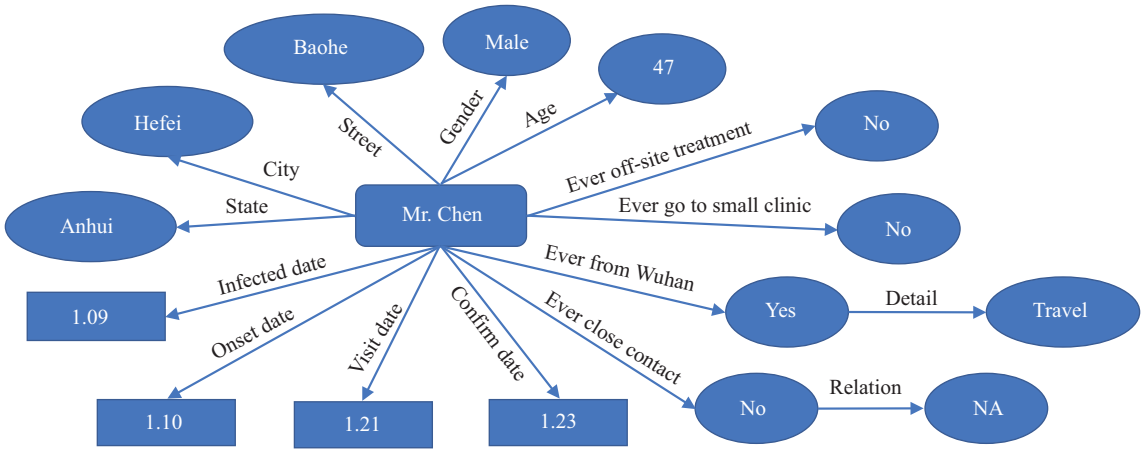


图 2 (网络版彩图) 结构化的患者病例详情示例

Figure 2 (Color online) An example of the structured Covid-19 case details

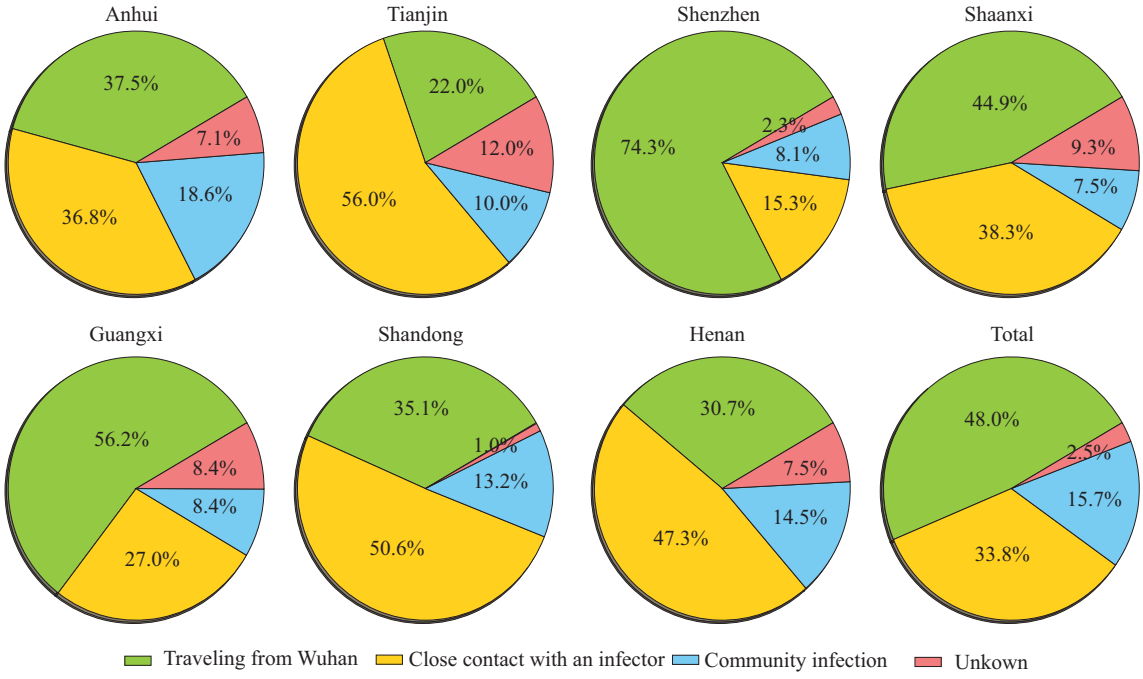


图 3 (网络版彩图) 感染原因

Figure 3 (Color online) Causes of infection

过 50%. 接触患者造成的感染包括家庭传染、同事传染、聚会传染、看望病人等形式的直接接触传染, 这也是疫情在各省市再次蔓延的重要原因. 本次疫情后期呈现家庭、朋友、同事等聚集性传染的特点. 在接触患者传染中, 家人亲属传染占据主要因素, 占比约 65%. 例如, 存在一些病例从武汉回来, 参加宴会, 导致家庭成员直接和间接感染十余例. 社区感染占比相对较少, 但一旦出现超过一定比例, 就说明一个地区疫情已经比较严重, 不可忽视.

如图 4, 截止 2 月 10 日, 具有武汉旅居史的病例占比总体呈现下降趋势, 接触患者造成的感染呈

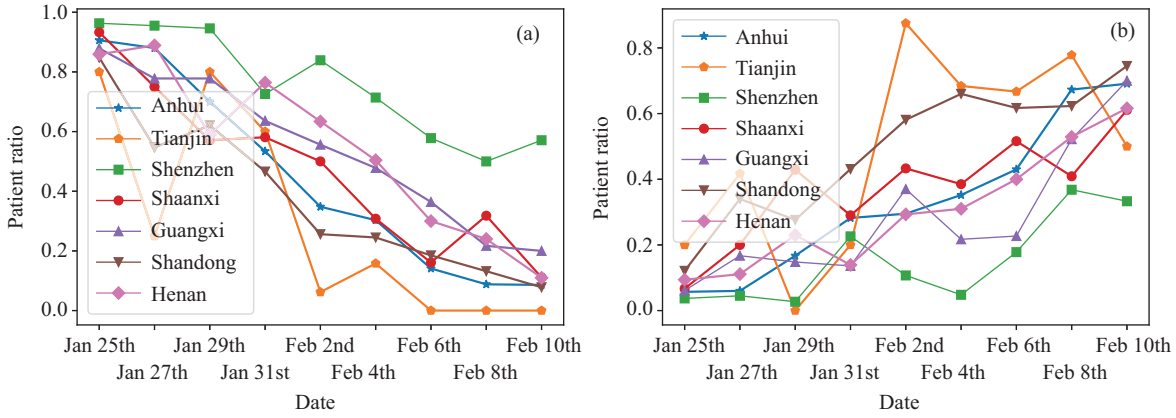


图 4 (网络版彩图) (a) “旅居武汉” 染病病例占比; (b) “接触患者” 染病病例占比

Figure 4 (Color online) (a) Proportion of infected “travellers to Wuhan”; (b) proportion of infected cases with “Close physical contact with an infector”

现爬升趋势. 从 2 月 11 日各省市的数据来看, 除深圳市外, 各省市基本上完成输入病例 (具有武汉旅居史的病例) 的就诊和确诊工作. 全国除湖北外的新增患者数据也呈现了大幅度下降趋势, 抗疫工作已初见成效.

4.2 性别比例

部分省市和总体性别比例如图 5(a), 可以看到男性患病人数比女性多了近 6%, 患者男女比例约为 1.14:1, 而我国自然人口男性比女性多 4.5%. 但数据显示, 虽然一开始男性比例相对较高, 随后男女比例越来越接近. 男性患者比女性患者多, 并不能轻易推断男性更易感, 这与一些社会因素也有关系. 2019 年腾讯发布的春节出行预测表明[7], 出行人群中, 男女比例接近 1.5:1, 而此次新型冠状病毒肺炎患病男女比例是 1.27:1. 所以 “男性比女性更易感” 这样最初的结论也是欠准确的, 忽略了出行人群的性别结构.

4.3 年龄分布

从图 5(b) 中可以看到, 患者的年龄分布大致服从一个以 44 岁为平均值的正态 (高斯) 分布, 且患者年龄分布和我国人口自然年龄分布一定程度上吻合, 说明并不存在不易感人群, 早期不易感人群的报道欠准确. 没有考虑到早期病毒爆发地点离福利院较近, 老人居住相对密集的地理特征. 本节我国人口自然年龄分布采用第 6 次人口普查自然年龄分布图, 虽然第 6 次普查到现在人口结构发生了一定变化, 65 岁以上人口增加了 2%, 15 岁以下人口减少了 0.2%, 但总体结构还是相对稳定的. 从患者年龄分布来看, 65% 以上患者占据 10%, 而我国自然人口年龄分布 65 岁以上人口占据总数的 11.4%. 从这个数据来看老人并不属于易感. 25 岁以下患者比例相对自然年龄比例相对较少, 但这并不等于不容易感染. 推测一开始感染病例务工人员居多, 孩子并没有带在身边. 另一个原因是年轻人接触病原体的机会相对较少, 更偏向或者需要在家里学习工作, 而造成了数据统计偏差. 30~65 岁年龄段, 感染人数比例大幅度超过人口自然年龄分布, 平均比人口自然年龄比例高出 3%.

7) https://new.qq.com/cmsn/20190125/20190125008337.html?pgv_ref=aio2015&ptlang=2052.

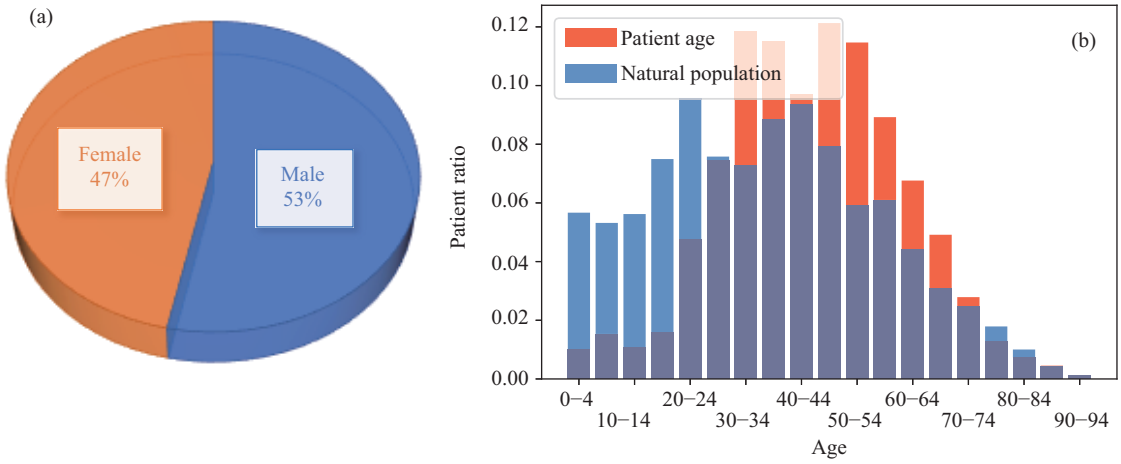


图 5　(网络版彩图) (a) 患者性别比例; (b) 患者年龄分布与我国人口自然年龄分布对比

Figure 5　(Color online) (a) Gender proportion of patients; (b) comparison of patient age distribution and China's natural popularity distribution

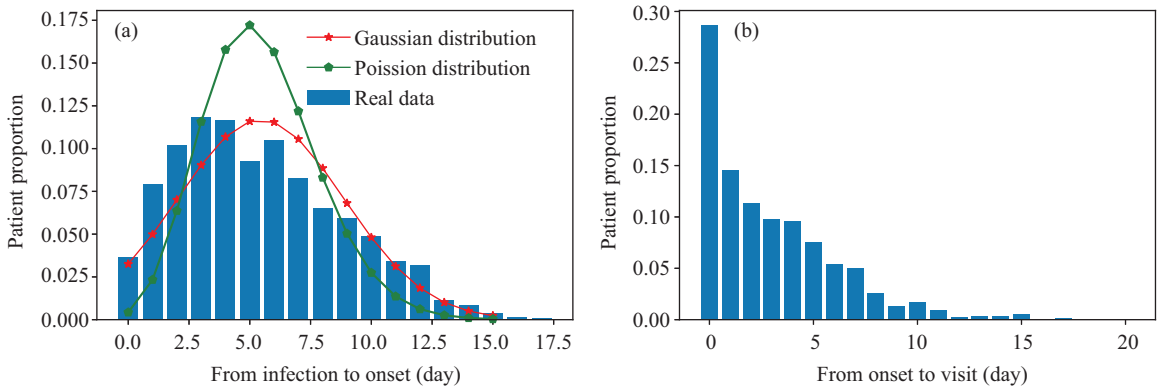


图 6　(网络版彩图) (a) 潜伏期的数据分布和拟合分布; (b) 就诊延迟的数据分布和拟合分布

Figure 6　(Color online) (a) Data distribution and fit distribution during the latent period; (b) data distribution and fit distribution of medical consultation delay

4.4　潜伏期分析

用图 3 中的发病时间减去感染时间计算患者发病潜伏期, 分布如图 6(a), 潜伏期平均时间 5.17 天, 中位数为 5.0 天. 大多在 2~7 天之间, 和中华人民共和国国家卫生健康委员会最初给出的 3~7 天潜伏期基本相符. 这和早期的统计结果 5.2 天 [8] 差别不大, 但和钟南山院士团队 [12] 的结果 4.0 天相差略大. 主要原因有数据样本不同, 本文数据不包括湖北省数据, 湖北省疫区病毒密度大, 可能潜伏期较短些, 也有可能病毒产生了变异或进化 [17]. 患者年龄分布不同, 在本文的数据样本中老年人比例低于钟南山院士团队的. 时间范围不同, 钟南山院士论文数据截止到 1 月 29 日, 本文数据截止到 2 月 10 日.

问题: 不断有报道说出现了潜伏期疑似超过 14 天的患者, 如何看待较长潜伏期 (超过 14 天) 的患者?

从图 6(a) 来看, 高斯分布比较符合此次的发病潜伏期分布. 在拟合高斯分布时, 我们采用最小二乘法计算高斯分布的参数 μ 和 σ. 我们给出在高斯分布下潜伏期超过 14 天的概率.

15 天: 1.0×10^{-3}; 16 天: 3.9×10^{-4}; 17 天: 1.3×10^{-4}; 18 天: 3.8×10^{-5}; 19 天: 1.0×10^{-5}; 20 天: 2.5×10^{-6}; 21 天: 5.42×10^{-7}; 22 天: 1.1×10^{-7}; 23 天: 1.1×10^{-8}; 24 天: 3.2×10^{-9}.

潜伏期 15 天出现的理论概率只有千分之一, 16 天的理论概率约是万分之四, 17 天的概率近万分之一, 20 天的概率为百万分之 2.5, 24 天的概率为十亿分之三. 但是考虑到这次传染面积大, 受众广, 小概率事件也存在一定的发生可能性.

潜伏期时间是概率事件, 影响因素非常复杂, 和病毒密度、患者体质、运动习惯都有关系. 从分布可以看到, 潜伏期并没有在 14 天出现截断的现象. 而是按照一定的概率出现, 天数越长, 概率越低. 但目前还没有确切公布的病例详情显示患者潜伏期超过 20 天. 我们对于超过 14 天潜伏期的病例, 会仔细核查病例信息, 并和地方卫生健康委员会联系核实病例, 多数是感染时间模糊 (患者去过很多地方, 难以推断真实感染时间)、笔误、统计工作失误造成的.

例如, 安徽合肥市有一例, 最初计算得到 23 天潜伏期存在记录笔误, 实际潜伏期只有 5 天. 目前只有数例 15 天以上潜伏期, 比如发生在安徽宿州市, 比较确信, 患者和妻子从武汉返回后住在一起一段时间, 无外出行为. 安徽亳州市一病例无武汉旅居史, 其女儿 1 月 5 日在重庆旅游, 1 月 26 日发病, 很难确定她的确切感染时间和感染方式, 只能说最早 1 月 6 日被女儿传染. 四川一例, 该患者早有症状, 因为服了感冒药, 缓解了后期的症状, 且患者后期又继续接触了病例, 并不能肯定潜伏期超 20 天. 以及其他疑似 15 天以上潜伏期病例大多存在多个时间接触感染源的现象. 理论概率也说明, 99.9% 的概率患者潜伏期是在 14 天以内. 14 天的隔离期, 也是考虑到事件发生概率、现实的发生率, 以及实际隔离条件作出的权衡. 不必因出现个例恐慌, 但要早发现早治疗. 超过 14 天之后, 事件发生的可能性成指数型快速衰减.

从生物生存与进化的角度来看, 潜伏期越长, 越有利于病毒与宿主共生, 不排除新型冠状病毒未来进化延迟期更长的可能性, 甚至未来会出现一个潜伏期更长的新病毒, 值得警惕.

4.5 就诊延迟分析

从图 6(b) 中可以看到, 仅有不到 30% 的人选择当天就诊, 超过 55% 的人群出现症状 2 天以后才去就诊, 平均发生症状后 3 天才会选择就诊, 从发病到就诊中位数时间是两天. 患者最长选择出现症状 20 天后才去就诊. 一些经济相对落后地区的病人, 发病后采用自行吃药的方式, 延迟甚至 10 天后才去就诊. 在此期间, 不仅会有更大机会感染给家人亲属, 甚至诊所和村医等医务人员, 还会造成病人本人错过最佳治疗时间转入重症. 有相当一部分病人是在自行治疗若干日之后, 通过 120 急救车送到医院的. 在疫情初期, 对落后的地区和思想观念落后的人群进行宣传教育非常重要. 早在 1 月 21 日疫情初期, 深圳市 1 月 21 日之前确诊的 15 例输入病例中, 11 例是异地就医. 在出现发热、咳嗽等症状的情况下, 不选择及时就医, 甚至外出, 增加了同行高铁、公交车乘客, 甚至乘务员感染的风险, 这也是新型冠状病毒为何短时间撒播到全国各地的一个原因.

4.6 确诊延迟分析

从图 7(a) 可以看到, 平均就诊到确诊的时间延迟最常见是 2 天, 平均 3 天, 中位数是 2 天. 见图 7(b), 确诊时间的延迟随就诊时间越来越小, 平均就诊到确诊时间所需天数从最初 3 天到 2 月 5 日的 0.74 天. 中位数从 3 天到 1 天. 在疫情初期改革了体制, 最初确诊需要县级医院和省级医院同时确认, 后面县级以上医院可以直接确诊. 1 月底改进了 RNA 核糖核酸检测技术, 到目前最快一次检测半小时内就可以出结果, 可见技术的进步对于抗疫的重要性. 但最初武汉市的试剂盒和检测技术人员都不足, 造成大量患者积压待检测, 因此建立一个 “预备役” 的特殊技术人才储备尤为重要.

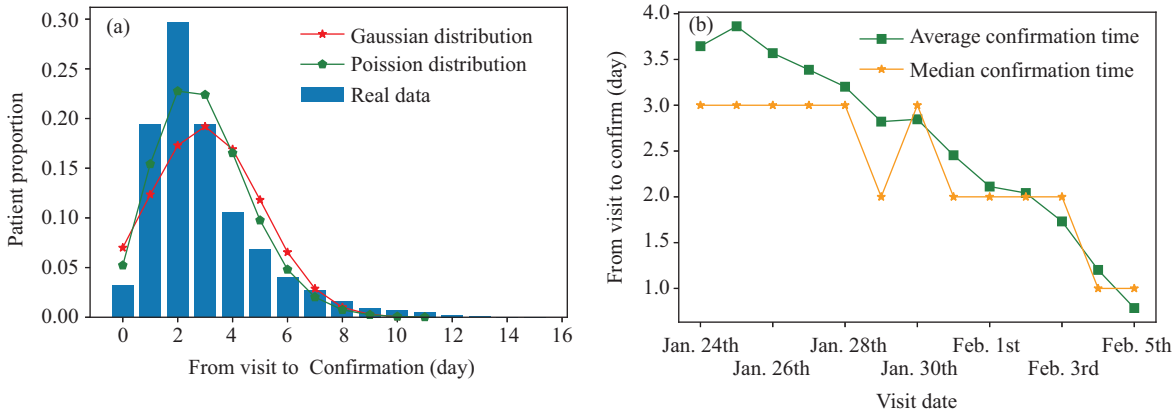


图 7 (网络版彩图) (a) 确诊延迟的数据分布和拟合分布; (b) 确认延迟随着就诊日期降低曲线

Figure 7 (Color online) (a) Data distribution and fit distribution of confirm delay after visiting. (b) Confirm delay reduces with visit dates

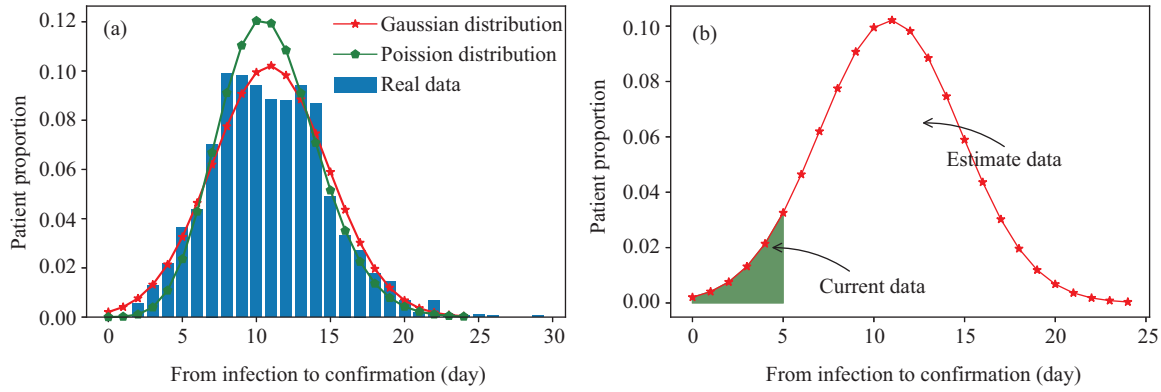


图 8 (网络版彩图) (a) 真实从感染到确诊时间数据分布和拟合分布; (b) 通过早期样本推测整体数据示例

Figure 8 (Color online) (a) Data of distribution and fit distribution from infection to confirmation; (b) an overall data estimation based on early samples

如图 8(a), 从发病到确诊, 平均有 10 余天的时间滞后. 这也是为何我们看到, 湖南省 1 月 23 日确诊了不到 24 例, 一周后却确诊了 300 多例感染者的原因.

4.7 基础再生数

基础再生数 (基础再感染人数), 是指一个患者平均再感染患者的数量. 如何计算基础再感染人数? 之前的文献大多基于小样本来估算武汉市封城前基础再感染人数 $R0$. 本文提出采用武汉市患者增长的大数据来计算基础感染人数. 如图 9 所示, 根据患者的感染时间的分布计算 $R0$. 如图 9, 武汉封城前, 患者数量的增长随时间呈现指数增长, 基本符合 $a \times b^n$ 的形式, a 是基础系数, b 是增长系数, 使用最小二乘法或者梯度下降方法求解可以得到 a, b. 求解得到 $b = 1.39$, 意味着每天病例增长 39%, 每个病例每天平均感染 0.39 个新病例. 即便考虑到春节返乡人流的增长, 依旧不改变指数的增长调性. 说明在封城前, 阻碍传染病指数增长的措施有限, 最快每隔 3 天感染人数可以翻一番. 如果再晚几天封城, 再晚几天采取严格的隔离措施, 后果不堪设想. 而潜伏期中位数 5 天, 就诊延迟平均 3 天. 不考虑就诊延迟, $R0 = 5 \times 0.39 \approx 2.0$. 考虑就诊延迟, $R0 = 8 \times 0.39 \approx 3.2$. 这只是考虑 1 月 23 日之前的武

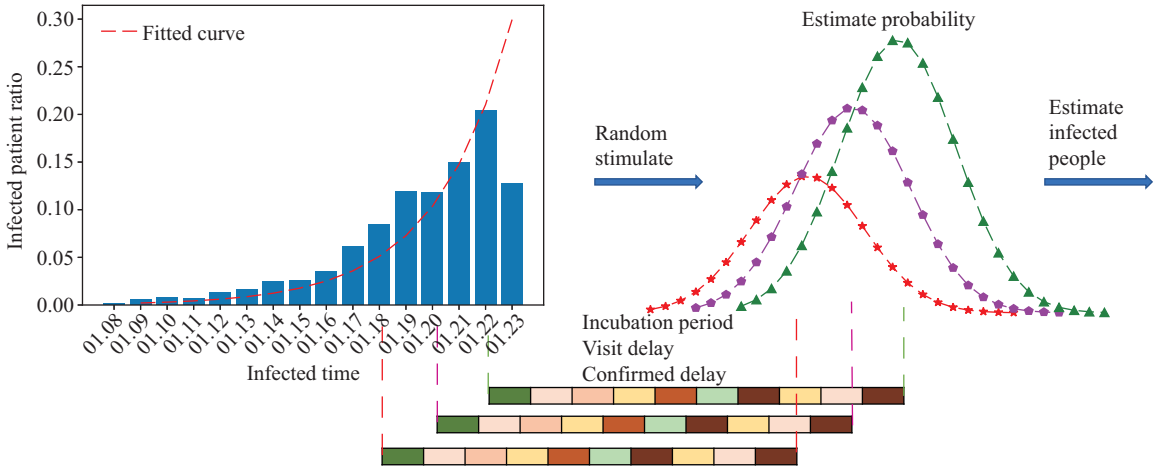


图 9 (网络版彩图) 通过早期样本推测整体数据

Figure 9 (Color online) An overall data estimation based on early samples

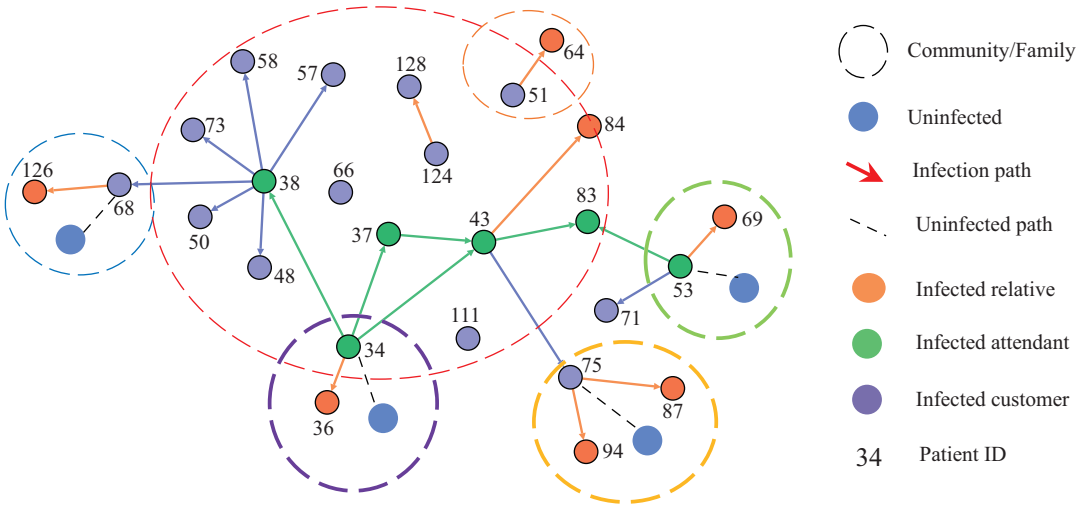


图 10 (网络版彩图) 天津宝坻区百货大楼病毒感染传播路径图

Figure 10 (Color online) Infection roadmap for the COVID-19 case occurring at a shopping mall in Baodi District, Tianjin

汉患者, 并非就全国数据得到的 $R0$. 除湖北省武汉市外, 其他省份、城市病原体传入时间较晚, 而且采取了积极的防控措施, $R0$ 不同. 当然, 随着各地措施的增强, $R0$ 在不断地降低, 最终接近零.

4.8 其他特点

从数据中发现, 存在大量患者跨区域就诊的现象. 共有一百余例患者存在异地就诊的情况. 1 月 20 日之后钟南山证实了存在人传染人的现象, 但依然有不少人选择隐瞒病情, 异地就诊. 深圳市情况最为严重, 截至 2 月 6 日, 存在超过 50 例异地就诊的病例. 同时有相当多的病例, 选择去村医、社区诊所就医, 导致村医和社区医生被感染, 造成更大范围的传染. 最初的数据就显著表明了存在大量异地就诊、社区就诊的现象, 却没引起足够重视.

社区连接者也称为结构洞人群 [18], 在各个社区之间负责 “信息联络”, 如图 10 所示天津市第 34

号、38 号、68 号和 75 号等病例. 从病例的职业关键词可以发现, 部分收银员、教师、乘务员、厨师、店员、服务员、医务人员等感染. 从流行病学和复杂网络的角度, 这是本次疫情爆发的一个关键原因. 大量社区连接者被感染, 甚至被不知情感染, 导致病毒从最初的小家庭传播到各个大社区之中. 一个显著的例子就是天津市, 我们提取天津市病例患者关系绘制图 10, 天津市输入病例只有 22 例, 截至 2 月 10 日, 却确诊了 100 例患者, 基础再生数 $R0$ 居全国前列. 其中, 一家超市工作人员 (34 号患者) 在北京采购被感染, 34 号传染给 38 号等同事, 随后造成更多感染, 近万人被隔离. 想要完全阻断各个传播路径比较难, 代价很大, 最有效的方法是控制 34 号等社区连接者不被感染, 控制这些节点, 造成级联失效, 能够大幅度降低传播的深度和广度 [19,20]. 控制和监控 “社区连接者” 的健康和活动, 尤其是服务较多客户的从业者不被感染, 把 34 号、38 号这样的结构洞人群控制住, 部分关键级联传播路径就被阻断. 图 10 是天津市百货大楼传播路径的真实例子. 控制人群买菜购物很难, 而控制收银员、服务员不被感染的代价则相对较小.

5 结合出行大数据估计武汉市感染人数

由于湖北省、武汉市发病早期医疗物资、医务人员、试剂测试能力等多方面的限制, 1 月 27 日之前试剂盒日检测能力为 200 份, 从 1 月 27 日之后提高到 2000 份, 但患病人数依然大量积压, 实际的感染人数远远超过每日确诊数目. 确诊数据无法反应武汉市真实感染情况, 这对了解武汉市实际感染人数、调动真实所需医务资源产生巨大的障碍. 但是我们依然可以通过出行大数据, 结合确诊病例结构化数据, 间接且合理地估计武汉市以及其他湖北省地级市的早期实际感染人数, 为疫情决策、物资调配、医院床位建设、人员调配提供数据参考. 本文以武汉市感染人数估计为例, 其他城市的感染人数也可以通过相同的方法进行估计.

5.1 方法描述

定义5 (人群感染率) 随机抽出某城市 100 人, 包含感染者的百分数.

一个城市的流入和流出人口数据是可以通过出行大数据得到. 假设武汉市 (城市 A) 的人口是充分混居的, 人群感染率为 P_{A}, 总感染人数为 N_{ga}, 武汉市总人口数量 N_{A}. 其中 $P_{\mathrm{A}} = (N_{\mathrm{ga}}/N_{\mathrm{A}})$. 武汉市, 约有 1059 万人常住人口, 500 万流动人口, 假设流动人口和居住人口充分混合. 从武汉市流向城市 B 的人口数为 N_{AB}, N_{AB} 里有 N_{gb} 人感染, 流向城市 B 的人群感染率为 P_{B}. N_{AB} 和 N_{gb} 相对易得到. 相当于武汉的盒子里红球白球若干, 红球的个数是 N_{ga}, 从一个叫做武汉的盒子里随机不放回抽出若干个球的问题, 是数学上组合问题. 抽到感染人群共有 N_{gb} 的概率: $C_{N_{\mathrm{A}}}^{N_{\mathrm{AB}}}(P_{\mathrm{A}})^{N_{\mathrm{gb}}}(1-P_{\mathrm{A}})^{N_{\mathrm{AB}}-N_{\mathrm{gb}}}$. 对于二项分布, 当 N 比较大时, 期望值是 $N \times p$, p 是事件发生概率. 一旦求解出 P_{A}, 就能求解出 N_{ga}, 反之亦然. 为了简化计算和容易阅读, 本文直接使用简化的方法.

已知武汉市春节期间离开约 500 万人, 流向信阳市的人口约有 1.4%, 计算得到 7 万人, 信阳市官方报道春节期间武汉市流入信阳市人口 7 万余人, 基本在一个量级. 武汉出行流动目的地城市示意图, 如图 11. 其中, 武汉市大约有 1.87 % 流向了深圳, 0.4% 的人口流向了合肥, 0.15% 流向了天津, 武汉流入不同城市数据, 以及 1 月 25 日、1 月 28 日以前的流入病例数据如表 2 所示. 注意本文旨在估计 1 月 23 日武汉市封城后 1059 万人口中的感染人数, 1 月 23 日后面的感染增长和封城后采取的防护措施严格程度有关.

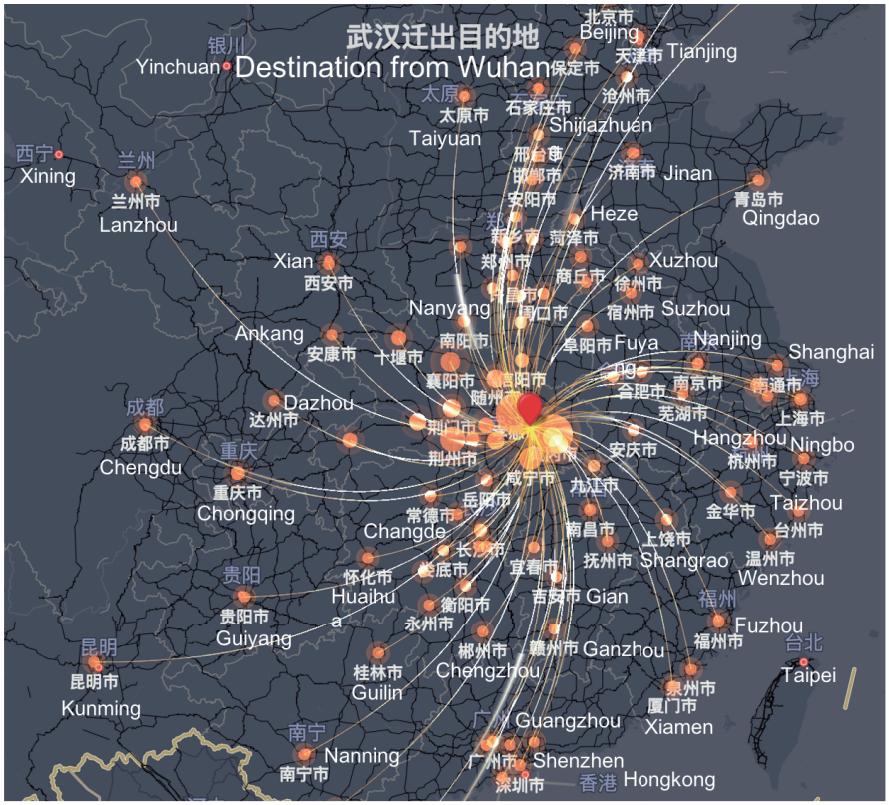


图 11 (网络版彩图) 武汉出行流动数据局部示意图 (来自百度出行 2020 年数据)

Figure 11 (Color online) Partial sketch of dynamic Wuhan traveling data (provided by Baidu Travel)

5.2 结果分析

当人群充分混合, 计算城市 A 的人群感染率 $P_{\rm A} \approx P_{\rm B} = N_{\rm gb}/N_{\rm AB}$, 根据 $P_{\rm A}$ 计算感染总数 $N_{\rm ga} = N_{\rm A} \times P_{\rm A}$. 由武汉市流入到不同的城市的病例数见表 2. 使用结构化的数据, 在确诊病例中, 只考虑武汉市 (城市 A) 向城市 B 的输入病例. 如果不使用结构化的数据, 则认为城市 B 的全部病例全部由于武汉市输入导致, 这是早期的感染估计模型大多采用的方法 [3, 4]. 根据 14 个城市的结构化病例详情数据或者非结构化病例估计武汉市感染人数如表 3 所示. 模型还要加上 1 月 23 日之前武汉市确诊的人数 549 人, 以及当时已经就诊和被隔离无法流出的患者, 至少两千人.

估计的结果汇总见表 3, 武汉市 1 月 23 日感染人数估计平均值为 3.23 万人, 中位数为 3.16 万人. 为了降低噪声, 可以去除一个最高值和一个最低值. 将全部城市的数据融合在一起, 当作一个城市进行估计得到武汉市感染人数 2.96 万人, 人群感染率约为 0.275%, 和 14 个城市分别估计的平均值和中位数接近. 当然少部分城市在 2 月 10 日后依然有武汉市输入病例存在, 但是其占比已经较低, 这个估计数略有低估.

如果采用非结构化数据估算, 例如, 以天津市确诊病例估算的感染人数, 武汉市感染人数为 11.42 万, 远高于当时可能的真实情况. 而结构化的数据估计则得到感染人数约为 3.16 万人, 与真实情况相对较为接近, 可以看到结构化病例数据在城市感染患者数量估计中的重要性. 综合估算数据, 武汉市在 1 月 23 日, 1059 万人中有 3 万余人感染, 人群感染率约 0.3%.

相比之下, 网络传言武汉市感染者 10 万人, 在大规模数据分析面前不攻自破. 以及通过日本包机

表 2　武汉市流入部分城市病例

Table 2　Cases floating from Wuhan to destination cities

City	Floating population rate (%)	Floating cases	Total cases	Cases before Jan. 25th	Cases before Jan. 28th
Hefei	0.4	41	104	9	16
Fuyang	0.35	59	105	10	19
Qingdao	0.12	19	43	6	10
Jinan	0.15	17	39	2	4
Heze	0.1	10	13	1	6
Tianjing	0.15	22	79	8	13
Shenzhen	1.87	261	334	26	61
Nanning	0.19	17	32	1	6
Guilin	0.13	22	28	10	15
Beihai	0.09	28	31	7	13
Zhumadian	0.66	74	107	6	13
Taizhou	0.54	76	124	21	39
Changde	0.33	67	93	10	31
Huaihua	0.11	17	38	5	10
Total	5.19	712	1170	122	256

表 3　根据不同城市数据估计武汉市人群感染率和感染人数

Table 3　The estimated the infection rate and infection cases in Wuhan based on data in different cities

City	Infection rate (%)	Infection cases ($\times 10^4$)	Infection rate (%, unstructured)	Infection cases ($\times 10^4$, unstructured)
Hefei	0.21	2.23	0.52	5.77
Fuyang	0.34	3.63	0.60	6.61
Qingdao	0.32	3.41	0.72	7.85
Jinan	0.23	2.46	0.52	5.77
Heze	0.18	1.96	0.26	3.01
Tianjing	0.29	3.16	1.05	11.42
Shenzhen	0.28	3.01	0.36	4.04
Nanning	0.18	1.95	0.34	3.83
Guilin	0.34	3.64	0.43	4.82
Beihai	0.62	6.65	0.69	7.56
Zhumadian	0.22	2.43	0.32	3.69
Taizhou	0.28	3.04	0.46	5.12
Changde	0.41	4.36	0.56	6.23
Huaihua	0.31	3.33	0.69	7.58
Average	0.30	3.23	0.54	5.95
Median	0.29	3.16	0.52	5.77

206 个乘客, 3 人感染, 估计人群感染率为 1.45%, 感染人数 15.4 万. 这种估计方法样本数量太少, 偏差大, 且没有考虑到在飞机封闭式环境下交叉感染的问题. 帝国理工大学的模型 [3], 以 1 月 24 日海外确诊人数 11 人 (考虑时区延迟一天), 估计武汉市 1 月 23 日的感染人数约为 6285 例. 美国东北大学 [4]

估计新型冠状病毒基本再生数 $R0$ 在 3.2∼3.9 之间, 武汉 1 月 23 日的感染人数约为 10540 例, 均和实际情况相差较大. 基于小数据和偏向性选择人群, 得到的结论与事实偏差较大. 我们采用更大规模的数据和更合理的模型建模问题, 可以在早期对一个城市的疫情进行合理估计, 既能充分重视, 又不引起恐慌.

基于不同城市的数据进行预测产生了一定的偏差, 因为模型没有考虑到: (1) 务工人员更容易产生聚集感染. 比如, 河南驻马店市、安徽阜阳市、广西北海市、浙江温州市等城市务工返乡人员感染率相对较高. (2) 返乡人员路上被感染的可能性. (3) 患者人员的社会层次. (4) 部分城市是交通枢纽, 部分患者中间目的地城市 (换乘, 中间城市探亲之类的), 后面可能还有最终目的地城市. (5) 从武汉市到目的地城市需要 0.5∼1 天的时间延迟.

5.3 更早预测

问题: 一定要等到 2 月份才能合理地估计武汉市感染人数吗, 有没有更早估计的合理方法?

答案: 有问号例如, 我们给出采用 1 月 28 日、1 月 25 日甚至 1 月 23 日的数据来估计武汉市感染人数的方法. 我们可以根据武汉市流向一个城市的人口, 及流入人口的感染率, 和使用之前求得从感染到确诊时间延迟的概率分布, 随机模拟计算未来流入到城市 B 的确诊人数随时间的分布, 可以得到一个类似图 8 的分布. 根据预测时间 T, 高斯分布的概率密度 1 和积累分布密度 2 推断未来流入城市 B 的病例总数, 间接地估计武汉市的感染人数.

在估计之前, 我们需要基于一定的历史数据, 计算患者从感染到确诊的概率分布和积累分布密度. 其分布相关数据早在 1 月 27 日高福院士团队完成的文献 [8] 中就已经可以得到. 实际分布情况可能比较复杂, 但样本较多时, 可以用概率分布模型大致拟合, 本文采用高斯分布进行拟合. 同理, 也可以采用泊松分布等进行拟合. 高斯分布概率密度函数为

$$f(x) = \frac{1}{\sqrt{2\pi}}\left(- \frac{(x-\mu)^2}{2\sigma^2}\right). \tag{1}$$

概率密度函数的积累分布函数:

$$F(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}}\int_0^x \left(- \frac{(x-\mu)^2}{2\sigma^2}\right)\mathrm{dx}. \tag{2}$$

我们就可以根据 T 时刻的 $F(T;\mu,\sigma)$ 估计最终的输入病例 $F(\infty;\mu,\sigma)$. 即在未知 N_{gb} 的情况下, 可以根据早期的确诊数据估计出 N_{gb}. 得到 N_{gb} 后, 再间接地计算 P_{A} 和 N_{ga}. 这种估计是一个大致推断, 而真实情况极度复杂, 很难用一个模型完美地还原. 选取多个城市的数据进行估计, 一些噪声会被大规模数据中和掉, 取中位数或者平均值可得到一个合理的估计数值.

本文采用两种方法, 实现在早期对一个城市 B 未来的流入病例总数 N_{gb} 的估计.

方法 1. 直接假设确诊数量随确诊时间的增长服从高斯分布. 这种方法相对简单, 如图 8(b) 所示, 我们根据早期观察的数据, 如绿色部分, 直接进行拟合估算整体的高斯分布参数 μ,σ, 计算得到 $P(T) = F(T;\mu,\sigma)/F(\infty;\mu,\sigma)$, 根据当前时间 T 的流入病例数, 计算未来流入病例的总人数 P_{gb}. 通过该方法, 我们可以计算得到 $P(T=23) = 0.059$, $P(T=25) = 0.14$, $P(T=28) = 0.359$. 将全部城市数据作为一个城市数据, 1 月 23 日流入病例数为 54, 1 月 25 日流入病例数为 122, 1 月 28 日流入病例数为 256, 以此分别估算武汉市封城前患病人数为 3.78 万人, 3.61 万人和 3.0 万人.

方法 2. 先模拟城市人流的活动再计算感染概率分布. 估计方法流程如算法 1. 这种方法模拟武汉市病例的增长和患者的流入流出, 再计算未来一段时间确诊患者的增长分布. 我们设 T_0 为 1 月 12 日

算法 1 The estimated infected case number in Wuhan based on early confirmed case data

Input: City B structured cases, date of city A start to export T_0, closing date T_{end}, left popularity N_{A} in city A, daily net outflow population N_F, proportion popularity α_T export to city B, the growth rate of infection cases b, Gaussian distribution μ and δ from infection to confirm, the predicted time T_p;

Ouput: Infection rate P_{A}, infection quantity N_{ga} of City A;

Current date $T \leftarrow T_0$, max confirm delay $\triangle T \leftarrow 20$ day, current infection quantity N_0, current net outflow population N_{F0};
while $T <= T_{\text{end}}$ **do**
  Outflow cases $C \leftarrow 0$;
  for $T_i = T$ to $T + \triangle T$ **do**
    According to μ, δ to calculate the probability of P_{T_i} at confirm dates T_i;
    Estimate the cases export to the city B with confirmed time T_i, $N_{T_i} = P_{T_i} \times N_0 \times \alpha_T \times N_{F0}/N_{\text{A}}$;
    Update confirmed cases in date T_i according to N_{T_i};
    Update outflow cases $C \mathrel{+}= N_{T_i}$;
  end for
  Update infection cases in city A: $N_0 = (1 + b) \times N_0 - C$;
  Update population in city A: $N_{\text{A}} \mathrel{-}= N_{F0}$;
end while
Calculate the probability P according to the distribution of the confirmed cases;
Calculate N_{gb} and P_{B} based on P and the structured confirmed cases of city B;
Calculate $P_{\text{A}} \leftarrow P_{\text{B}}$ and $N_{\text{ga}} = N_{\text{A}} \times P_{\text{A}}$.

(1 月 12 日之前病例数量较少, 很少流出), T_{end} 为 1 月 23 日, 根据模拟的结果, 得到 1 月 25 日积累分布的比例占整体的比例约为 0.158. 以合肥市为例, 根据表 2 中合肥市 1 月 25 日的输入病例数 9 例, 可以推测未来合肥市输入病例总数 N_{gb} 为 57 例. 再根据 N_{gb} 计算 P_{B} 和 P_{A}, 根据 P_{A} 计算 N_{ga} 为 3.1 万人. 由于初期每个城市确诊数据较少, 不确定性较大. 我们融合全部 14 个城市数据当作一个城市来估计武汉市感染人数, 根据 1 月 25 日和 1 月 28 日数据估计分别为 3.23 万人和 2.74 万人. 使用 1 月 25 日和 1 月 28 日之前 14 个城市的数据分别估计武汉市感染人数, 得到中位数分别是 3.98 万人和 3.56 万人, 预测的结果虽然有浮动, 但和真实情况的量级基本相符合.

模型还可以考虑更加复杂的因素, 例如, 考虑中国城市交通网络, 目的城市向其他城市人口的再流出. 我们认为城市人口短期内比较稳定, 流出的患者和流入的患者数量基本相当, 以简化模型. 此外, 返乡人员存在交叉感染的现象, 有些病例并不一定是在武汉时被感染, 而可能是在返乡的路上被感染. 比如, 北海市, 只有不足 0.1% 的返乡人员, 却存在 28 例输入病例. 研究发现, 以劳务输出为主的城市, 感染率相对较高, 见表 3. 劳务人员更容易聚集和一同往返, 部分劳务人员缺少自我保护意识, 路途上被感染的概率相对较大. 在武汉市 1 月 23 日确诊 549 例时, 很难想象武汉市已经有 3 万余人感染, 只有不到 2% 的患者被确诊, 真实感染人数是确诊人数的 50 倍, 大量病例积压未确诊, 这背后有复杂的原因, 而结合出行大数据和结构化数据, 可以推测数据后面隐藏的真相. 基于大量数据建模, 简易的模型却能得到一个量级相当的估计.

每年都会有新的病毒产生和旧的病毒 (例如, 流感病毒) 重现, 类似的方法经过扩展, 可以及早地估计一个地区人群感染率和感染人数. 并以此为基础, 结合大数据信息系统, 对一个地区的疫情及时分析和预警, 及时配备医务人员和医疗资源, 防止疫情大规模扩散.

6 总结与展望

本文在自然语言处理技术的辅助下, 对部分省市新型冠状病毒病例的文本数据进行结构化. 基于结构化病例数据, 对患者年龄、性别和感染原因, 以及病毒潜伏期、就诊确诊延迟、基本再生数等特征进行详细分析, 指出潜伏期等特征符合的分布特征. 在地区实际感染人数的估计上, 提供了一个相对合理的方法. 研究表明, 借助人工智能技术, 可以辅助病例数据结构化分析, 挖掘更多数据潜藏的真相. 在数据样本较少的情况下, 通过其他领域的大数据, 科学地建模, 能够较为合理地估计一个地区的感染人数, 及早地进行决策和战略部署. 本文所提研究方法也可以应用于海外疫情的分析之中. 例如, 美国 Dr. Fauci 预测美国将会有百万人感染, 采用的就是基于高斯分布的评估方法[8)]. 同样, 意大利也可以根据出行数据和早期流入到其他欧洲国家的病例数据估计意大利的真实感染人数. 该方法也可以用于估计其他欧洲国家的真实感染数据.

未来我们将汇集更大规模的数据, 矫正数据偏差, 并提供更多维度的数据分析. 以及在结合出行大数据的同时, 考虑城市间的流入流出. 大规模数据分析还给我们一些重要启示: (1) 建立地方和国家疫情信息的共享联动, 第一时间知晓和把控疫情信息, 尤其对于发热、肺炎、肝病、腹泻等门诊和住院的病例进行实时信息传输, 对数据建模监控预警, 防止因地方瞒报而产生误判. (2) 由于潜伏期、就诊延迟、确诊技术等影响, 早期确诊病例数据会诱导对疫情形势产生严重误判. (3) 通过大数据的方法和数学建模, 及时估计潜在感染人数, 能够及早估计到问题的真实情况和严重程度, 尽早采取措施. 而等到病例全部确诊时, 中间延误时间太久, 就会贻误时机. (4) 加强基层社区和村落的疫情防控, 对无知无畏的部分潜在传染源进行教育劝阻, 阻断异地就诊、隐瞒病情, 防止出现一人得病, 全家感染甚至整个社区隔离的现象. (5) 重视领域专业知识和专业技能整合, 将先进信息技术和交叉学科技术合理应用在疫情防控中.

致谢 感谢安徽省卫生健康委员会、安徽省共青团、广西省卫生健康委员会、湖南省卫生健康委员会、江苏省湖南省卫生健康委员会、山东省卫生健康委员会、陕西省卫生健康委员会、深圳市卫生健康委员会、天津市卫生健康委员会等提供的匿名病例数据支持, 感谢部分地级市卫生健康委员会对数据确认的支持, 感谢百度出行提供出行大数据支持. 感谢匿名评审专家耐心细致的指导. 特别致敬奋战一线的医务人员和志愿者!

参考文献

1 Li W. Research on key algorithms of mining texts of electronic medical cases. Shenyang: Northeastern University, 2014

2 Lu S Q, Dou Z C, Wen J R. Research on structured data extraction in surgical cases. Chin J Comput, 2019, 42: 2754–2768

3 Imai N, Dorigatti I, Cori A, et al. Estimating the potential total number of novel Coronavirus cases in Wuhan City, China. Imperial College London, 2020. https://www.imperial.ac.uk/media/imperial-college/medicine/sph/ide/gida-fellowships/Imperial-College-COVID19-epidemic-size-17-01-2020.pdf

4 Chinazzi M, Davis J T, Corrado G, et al. Preliminary assessment of the international spreading risk associated with the 2019 novel coronavirus (2019-nCoV) outbreak in Wuhan city. 2020. https://www.apprise.org.au/wp-content/uploads/2020/01/Chinazzi-CIDID20_nCoVExportation.pdf

8) https://mil.news.sina.com.cn/dgby/2020-04-01/doc-iimxyqwa4445974.shtml.

5 Wu F, Zhao S, Yu B, et al. A new coronavirus associated with human respiratory disease in China. Nature, 2020, 579: 265–269
6 Zhou P, Yang X L, Xian G, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature, 2020, 3: 1–4
7 Xu X, Chen P, Wang J, et al. Evolution of the novel coronavirus from the ongoing Wuhan outbreak and modeling of its spike protein for risk of human transmission. Sci China Life Sci, 2020, 63: 457–460
8 Li Q, Guan X, Wu P, et al. Early transmission dynamics in Wuhan, China, of novel coronavirus-infected pneumonia. New Engl J Med, 2020, 382: 1199–1207
9 Huang C, Wang Y, Li X, et al. Clinical features of patients with 2019 novel coronavirus in Wuhan, China. Lancet, 2020, 395: 497–506
10 Holshue M L, DeBolt C, Lindquist S, et al. First case of 2019 novel coronavirus in the United States. New England J Med, 2020, 382: 929–936
11 Chen N, Zhou M, Dong X, et al. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. Lancet, 2020, 395: 507–513
12 Guan W-J, Ni Z-Y, Hu Y, et al. Clinical characteristics of coronavirus disease 2019 in China. New Engl J Med, 2020. doi: 10.1056/NEJMoa2002032
13 Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017. 6000–6010
14 He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016. 770–778
15 Devlin J, Chang M, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019. 4171–4186
16 Ma X Z, Eduard H. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 2016. 1064–1074
17 Tang X, Wu C, Li X, et al. On the origin and continuing evolution of SARS-CoV-2. Natl Sci Rev, 2020. doi: 10.1093/nsr/nwaa036
18 Fang B X. Online Social Network Analysis. Beijing: Electronic Industry Press, 2014
19 Huang Z, Wang Z, Zhu Y, et al. Prediction of cascade structure and outbreaks recurrence in microblogs. In: Proceedings of Chinese National Conference on Social Media Processing, 2017. 53–64
20 Xu X K, Hu H B, Zhang L, et al. Computational Communication on Social Networks. Beijing: Higher Education Press, 2015

Analysis of COVID-19 spread characteristics and infection numbers based on large-scale structured case data

Zhenhua HUANG[1,2*], Zhenyu WANG[1*], Li JIANG[3], Rui ZHANG[1], Chang LEI[1], Xingwei LIU[2] & Xiaohui XIE[2]

1. *School of Software Engineering, South China University of Technology, Guangzhou* 510006*, China*;
2. *Department of Information and Computer Science, University of California Irvine, Irvine* 92617*, USA*;
3. *Department of Pharmacy, The First Affiliated Hospital of Anhui Medical University, Hefei* 230032*, China*
* Corresponding author. E-mail: sezhhuang@mail.scut.edu.cn, wangzy@scut.edu.cn

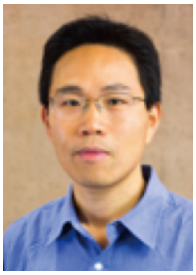
Abstract In early 2020, the novel coronavirus, referred to as COVID-19 burst out. The Chinese people took the most comprehensive and rigorous control measures to fight against the COVID-19. Local health control departments reported infection data in a timely manner, which helped the public understand the development of the epidemic and take protective measures in advance. However, currently, no literature has analyzed the transmission characteristics of COVID-19 based on the structured data of large-scale patient cases and artificial intelligence. The detailed case data of patients in various regions are primarily recorded in text form, and the formats of report data in different provinces and cities differ, which makes it difficult to handle such data. To analysis around a large anonymous patient case data, we propose a method based on natural language processing technology to structure the case data. The proposed method can extract key information in the cases accurately and effectively with the help a pretrained model and a small number of labeled samples. By mining the patient's structured case data, we analyze the gender and age distribution, the main causes of infection, the characteristics of the incubation period, and epidemic trends in detail. Using big data on travel, a method was developed to estimate the number of infected individuals in Wuhan prior the restrictions were put into effect. This method helps people understand the real epidemic situation and take execute early protective measures. It is also helps government departments make evidence-based decisions, dispatch medical staff, and allocate medical resources as quickly as possible.

Keywords coronavirus, structured medical cases, natural language processing, pretrained models, COVID-19 transmission characteristics, big data of traveling

Zhenhua HUANG was born in Anhui, China. He is currently a Ph.D. candidate in the School of Software Engineering at South China University of Technology, Guangzhou and the University of California, Irvine. Supported by the IBM-CSC Y-1000 young big data scientist plan, he is currently working at University of California, Irvine. His research interests include social computing, text analysis, and deep learning.

Rui ZHANG received his B.S. degree from the School of Software Engineering at the South China University of Technology, Guangzhou. He is currently a Ph.D. candidate in the School of Software Engineering at South China University of Technology. His research interests include natural language generation, text mining, and sentiment analysis.

Zhenyu WANG received his Ph.D. degree from Harbin Institute of Technology in 1993. He is a professor and the director of the Chinese Information Community of China. He is also the director of the Guangdong Provincial Social Media Processing and Engineering Center. His research interests include natural language processing, text mining, and social network analysis.



Xiaohui XIE is a Ph.D. student at the Massachusetts Institute of Technology, a postdoc at Harvard University, a professor in the Department of Computer Science at the University of California, Irvine. His research interests include machine learning, neural networks, deep learning on medical images, and genomics.