·论著·

〔文章编号〕 1007-0893(2022)22-0001-04 DOI：10.16458/j.cnki.1007-0893.2022.22.001

基于 Python 对深圳市新冠病毒感染者住院诊疗数据的挖掘研究

唐劲松 黄 华 毛 婷 田一梅 廖雪姣 邓 欣※

（深圳市第三人民医院，广东 深圳 518112）

〔摘 要〕 **目的：**通过对深圳市新冠病毒感染者诊疗数据的分析，比较 2020 年 1 月至 2022 年 3 月新冠病毒感染者临床特征，寻找新冠病毒感染者病情变化规律。**方法：**采用 Python 程序设计工具，对专病数据预处理，筛选出 1668 条有效治疗记录数据分析，采用 χ^2 检验比较治疗数据以寻找其差异变化，K–means 聚类模型分析，并可视化直观展示，比较不同新冠病毒株临床表现，归纳深圳市 2022 年以来新冠病毒奥密克戎毒株的致病特点。**结果：**深圳 2022 年新冠病毒奥密克戎 BA.2 株传播速度快，但毒性弱，重症率低，复阳人员总体病情较轻，较 2020 年经典新冠病毒株复阳率低，但平均复阳间隔天数缩短。**结论：**定期采用数据挖掘手段对新冠病毒感染者诊疗数据追踪分析，可发现不同时期、不同新冠病毒毒株感染者病情演变规律，为新冠病毒疫情防控提供科学的决策依据。此外少年感染人数及比率显著上升，其中愈后少年复阳者大多在 1 周内二次感染，推断这段时间是易感期，要加强宣传提高安全意识，注意病毒防护。

〔关键词〕 新冠病毒；新冠病毒感染者；Python；深圳市

〔中图分类号〕 R 563.1 〔文献标识码〕 B

Research on the Diagnosis and Treatment Date of COVID-19 Patients in Shenzhen Based on Python

TANG Jin-song, HUANG Hua, MAO Ting, TIAN Yi-mei, LIAO Xue-jiao, DENG Xin※

(Shenzhen Third People's Hospital, Guangdong Shenzhen 518112)

〔**Abstract**〕 Objective Through the analysis of the diagnosis and treatment data of coronavirus disease 2019 (COVID-19) infected people in Shenzhen, the clinical characteristics of COVID-19 patients from January 2020 to March 2022 were compared to find out the changing rules of the disease conditions of COVID-19 patients. Methods The Python programming tool was used to preprocess the data of specific diseases, and 1,668 effective treatment records were selected for analysis. The χ^2 test method was used to compare the treatment data to find the differences. The K-means clustering model was analyzed and visually displayed to compare the clinical manifestations of different novel coronavirus strains. To summarize the pathogenic characteristics of Omicron strain of COVID-19 in Shenzhen since 2022. Results The transmission rate of Shenzhen 2022 COVID-19 Omicron BA.2 was fast, but the virulence was weak, the rate of severe illness was low, and the overall condition of the patients who relapse positive was relatively mild, which was lower than that of the classic COVID-19 strain in 2020, but the average number of days between relapse positive was shortened. Conclusion Regular data mining is used to track and analyze the diagnosis and treatment data of COVID-19 patients, which can discover the evolving rules of the disease of patients with different COVID-19 strains in different periods, providing scientific decision-making basis for the prevention and control of COVID-19. In addition, the number and rate of adolescent infection increased significantly, and most of the adolescents who recovered from the disease were infected again within 1 week, suggesting that this period is a susceptible period. Therefore, it is necessary to strengthen the publicity, improve the awareness of safety, and pay attention to virus protection.

〔**Keywords**〕 COVID-19; People infected with COVID-19; Python language; Shenzhen

自我国爆发新冠病毒感染疫情以来，患者中的重型及危重型所占比例较高，有一定死亡率[1-2]。随着时间的推移，新冠病毒毒株不断演化，目前，我国主要流行的是新冠病毒奥密克戎变种，和以前的毒株比较，一方面

〔收稿日期〕 2022 - 09 - 22

〔基金项目〕 深圳市科技抗疫专项项目（技术攻关类一般项目）（JSGG20220226090002003）

〔作者简介〕 唐劲松，男，系统分析师（副高），主要研究方向是医院信息化管理及数据挖掘与分析。

〔※ 通信作者〕 邓欣（E-mail：szdengxin@126.com；Tel：13500054723）

新毒株毒性降低，死亡率大幅下降；另一方面传播力却增强了很多，防感染策略代价增大。深圳市第三人民医院是深圳市定点新冠病毒感染者收治医院。本研究收集了深圳市第三人民医院自 2020 年 1 月以来的新冠病毒感染住院患者的基础信息及诊疗数据情况，对其数据情况进行统计分析，展示新冠病毒感染患者的人次、住院天数、是否复阳、是否重型、复阳间隔等属性特征情况，探讨深圳市 2022 年以来新冠病毒奥密克戎毒株传染性和致病性的变化。

1 资料与方法

1.1 数据资料

数据取自深圳市第三人民医院的新冠病毒感染住院患者数据，主要有两部分数据，患者基本情况：住院流水号、住院号、年龄、国籍、籍贯、入院时间、主诉、现病史、流行病学史、出院诊断、是否使用呼吸机、是否使用重症监护室（intensive care unit，ICU）、出院时间、是否复阳、分型等，以及患者的病情变化、病例分型、转归、加重原因等信息。范围在入院时间 2020 年 1 月 1 日至 2022 年 3 月 31 日，数据直接导出为 excel 文件，初始数据共 2617 条。

1.2 数据挖掘方法

采用 Python 语言 numpy、pandas 库通过程序设计进行批量数据处理，方法与步骤参见表 1，最后可用的数据集共 1668 条。

表 1 数据预处理方法与步骤

序号	处理方法	数据 / 条
1	数据导出，按照住院流水号进行同一次住院数据的横向合并	2617
2	入院时间与出院时间比对检查，生成新列“入院年月”，并计算“住院天数”，个别空的出院时间采用平均住院天数与入院时间相加填充	2617
3	设置【年龄段】，年龄数字化处理并按照国际通用标准：少年 0 ~ 14 岁、壮年 15 ~ 64 岁、老年≥ 65 岁划分	2617
4	设置【重型】标志：a. 选择病例分型中所有“重型、危重型”的记录；b. 在“是否使用呼吸机”“是否使用 ICU”中的确认患者；c. 在“出院诊断”中写明“高危”的患者	2617
5	对【籍贯】省份的规范处理，只保留省份名称信息	2617
6	对【是否复阳】检查刷新：a. 住院数据按入院时间及住院号排序，对于每个患者（相同住院号）除了首次住院，随后的所有入院记录均标识复阳，并自动计算与上次出院时间的复阳间隔天数；b. 在“主诉，现病史，出院诊断”中搜索“复阳”文字并标识	2617
7	数据空值检查，对个别缺失比例超出 10 % 的数据列自动删除，如疫苗接种次数因缺失太多而被删除；个别记录因关键内容缺失而删除；个别数值栏采用均值填补法进行缺失值处理	2507
8	【分型】：发现数据中从 2022–02–08 开始标有“奥密克戎 BA.2”，除该毒株外其它记录全部为空或者未分型，在“现病史”中搜索“奥密克戎”最早 2022–01–16 就有该诊断，结合奥密克戎最早于 2021 年 12 月 9 日在天津市境外输入发现，这里设定数据从 2022–01–01 起的入院患者【分型】全部为“奥密克戎 BA.2”	2507
	数据检查	
9	从全部数据中检查【出院诊断】，检索不含字符串“新型冠状病毒”，找出 45 条，逐条通过患者“主诉、现病史、入院诊断、出院诊断”信息核实后删除了其中 42 条非新冠患者记录	2465
10	检查复阳记录，发现其中有数百条记录的复阳间隔时间小于 1 d，经仔细检查【现病史】内容，发现“因患者无居家健康监测条件，故再次办理入院，原地进行健康监测”，随后又陆续多次发现“现予以本院居家监测医学观察”“患者已达到出院标准”“今日予办理出院”“今日入院转居家隔离”“现患者转入健康观察阶段”等关键词的非正式入院记录，合计删除共 490 条假复阳入院记录	1975
11	仍然有不少复阳间隔时间为 0 天的情形，原因各异：医院应急院区增扩而转移病床、病危转入 ICU、医院信息系统的医保结算从市平台切换国家平台、其它临时出院再入院等情形，统计删除记录 307 条，并合并住院时间	1668

对 1668 条新冠数据采用数据驱动方式进行挖掘，即根据数据找问题、找合适的挖掘方法 [3]。运用 Python 语言的 Sklearn、Scipy、Matplotlib、Seaborn 等库进行数据挖掘与结果展示。因所有数据均没有死亡病例，本文作者初步考虑以“重型”作为标签，选择出一些有意义的特征属性：住院人次、住院天数、出院诊断、年龄段、是否复阳、复阳间隔等进行分析。

1.2.1 新冠病毒感染者籍贯的中国区域分布 首先对患者籍贯进行检索统计，排除了 210 人籍贯不详以及外国人合计 230 人，得出新冠住院患者籍贯的中国区域分布数据。

1.2.2 新冠病毒感染重型复阳患者关联规则 分析 2020–2021 年新冠病毒感染复阳患者疾病诊断数据，对个别疾病诊断合并同类项，如高血压 2 级、3 级合并，新冠病毒感染轻型与轻型恢复期合并等。采用关联规则 apriori 算法进行出院诊断的数据挖掘，其中频繁项集的支持度（support）15 %、规则的置信度（confidence）95 %，数据挖掘后得到频繁项集及其满足条件关联清单，展示为新冠重型复阳患者关联规则网络图。

1.2.3 新冠病毒感染者年龄分布 对 1668 条数据进行统计比较，先后绘制出了新冠患者各年龄段住院人次统计图、新冠患者 2022 年龄分布图。

1.3 统计学方法

运用 Python 语言的 Sklearn、Scipy、Matplotlib、Seaborn 等库进行数据挖掘与统计分析。采用 Scipy.stats 库的 χ^2 检验，对复阳数据进行前后时段的相关差异分析，$P < 0.05$

为差异具有统计学意义。选取年龄段、住院天数、复阳间隔作为分析属性，进行 K–means 簇类分析绘制新冠病毒感染者簇类特征散点分布图。

2 结 果

2.1 新冠病毒感染者籍贯的中国区域分布

提取患者籍贯中的省份地区并统计，除了西藏自治区外全国各省均有分布，其中患者数量排名前五的省份地区是：湖北 357 人、广东 264 人、香港 246 人、湖南 82 人、河南 65 人，正对应了国内先后的两大新冠热点地区湖北及广东香港。

2.2 新冠病毒感染患者重型复阳关联规则

2020–2021 年新冠病毒感染患者重型复阳关联规则网络图（支持度 15 %，置信度 95 %）见封三图 1。按照频繁项集的支持度排序，最高两项为高血压（45 %）与糖尿病（35 %），无论老年还是中年均属于易感人群；如果支持度下调到 10 %，则网络图更复杂、显现更多基础性疾病，有脂肪肝、高脂血症、腔隙性脑梗死、冠状动脉粥样硬化性心脏病等。

2.3 新冠病毒感染者重型与复阳人次

从封三图 2 中可以发现，新冠病毒感染住院患者的主要峰值时间分布在前后两大时段：一个是 2020 年 1 月至 3 月，一个在 2022 年 1 月至 3 月，前者对应当时武汉新冠病毒感染在我国初次爆发扩散，后者起因于香港今年初奥密克戎病毒疫情蔓延对深圳本地的冲击。重型患者总数很少，近三年全部重型患者 109 例，除了 2020 年初期开始爆发有一个小峰值，此后基本上单月都维持在个位数 4 例以下，即使在 2022 年 2 月新冠总数已达 602 例的高峰重型也仅 7 例。这说明奥密克戎新冠毒株虽然传播速度快，但危害程度已有明显下降。

2.4 新冠病毒感染者年龄分布

新冠病毒感染者各年龄段住院人次统计见封三图 3A，新冠病毒感染者 2022 年 1 月至 3 月年龄分布见封三图 3B。从图 3 中可以看到，2022 年 1 月至 3 月 0 ~ 14 岁少年的感染数量和比率显著上升，3 月份少年住院人次占比达 26.1 %（35/134）。

2.5 新冠病毒感染者复阳数据的差异性分析

2022 年新冠病毒感染者的复阳率较 2020–2021 年显著下降，其中主要为少年、壮年的复阳率下降显著，差异均具有统计学意义（$P < 0.001$）；2022 年新冠病毒感染者的少年人次占比相比于 2020–2021 年有显著增加，差异具有统计学意义（$P < 0.001$），见表 2。

复阳者的平均住院天数远远小于首次住院天数，特别是 2022 年以来复阳者住院时间更短，说明复阳病情较轻，很快就可以核酸转阴出院，2022 年的平均复阳间隔天数比 2020–2021 年缩短了一半多时间，见表 3。

表 2 新冠病毒感染者不同时间段的复阳人次、复阳率比较

时间段	年龄段	首次 / 人次	复阳 / 人次	总数 / 人次	人次占比 /%	复阳率 /%
2020–2021 年	少年	24	8	32	3.7	25.0
	壮年	642	137	779	89.2	17.6
	老年	53	9	62	7.1	14.5
	合计	719	154	873	–	17.6
2022 年	少年	140	3	143	18.0[a]	2.1[a]
	壮年	591	24	615	77.4	3.9[a]
	老年	36	1	37	4.6	2.7
	合计	767	28	795	–	3.5[a]

注：与 2020–2021 年同年龄段比较，[a]$P < 0.001$。

表 3 不同年龄段的新冠病毒感染者在不同时间段的住院、复阳间隔时间比较 （$\bar{x} \pm s$）

时间段	时间指标 /d	n	少年	n	壮年	n	老年	n	合计
2020–2021 年	首次住院天数	24	19.5 ± 6.4	642	21.8 ± 9.8	53	25.2 ± 12.2	719	22.0 ± 10.0
	复阳住院天数	8	7.9 ± 5.3	137	13.2 ± 9.9	9	10.1 ± 8.4	154	12.7 ± 9.7
	复阳间隔	8	13.5 ± 9.7	137	26.4 ± 73.9	9	17.0 ± 20.1	154	25.2 ± 69.9
2022 年	首次住院天数	140	19.2 ± 5.6	591	21.9 ± 5.2	36	23.5 ± 7.7	767	21.5 ± 5.5
	复阳住院天数	3	4.3 ± 4.9	24	9.1 ± 6.4	1	3.0 ± 0.0	28	8.4 ± 6.3
	复阳间隔	3	6.7 ± 5.7	24	11.4 ± 11.1	1	23.0 ± 0.0	28	11.3 ± 10.8

2.6 新冠病毒感染者簇类特征散点分布

将年龄段（少年 / 壮年 / 老年）数值化处理分别对应 1/2/3 簇，可以看到，复阳患者（复阳间隔大于 0 d）绝大部分都是出院不久后的二次感染，只有第 2 类（壮年）几例复阳间隔时间超出 90 d，见封三图 4。

3 讨 论

综上所述，将不同新冠病毒毒株阿尔法、贝塔、德尔塔到变种奥密克戎 BA.2 感染后的致病特点比较研究，对新冠疫情防控具有一定价值[4-5]。分析 2020 年 1 月至 2022 年 3 月医院新冠病毒感染患者住院数据发现，在今年初深圳流行的奥密克戎 BA.2 毒株的传播特性较以往已经发生了一些改变：（1）奥密克戎传染性强，传播速度快，但毒性弱，重症率低，对身体危害程度较低；（2）新冠复阳率有显著下降（从 17.6 % 降至 3.5 %，幅度达 80 %）；（3）复阳人员总体病情较轻，住院时间较短，复阳者的平均住院天数小于 9 d；（4）0 ~ 14 岁少年感染人数及比例显著上升；（5）平均复阳间隔天数缩短，特别是少年

的平均复阳间隔时间小于 7 d，在新冠治愈后，短期内要提高安全意识，注意新冠病毒防护，避免二次感染。本文作者分析深圳市新冠疫情研究结果，发现与吕莹等[6]报道的上海境外输入奥密克戎变异株感染患者的临床特点稍有不同，主要表现在 0～14 岁少年感染人数及比例较高。

按照当前国家的动态清零政策，我国有效地控制了新冠疫情，与国外与病毒共存的结果有天壤之别[7-8]。研究数据显示，2022 年 4 月以前深圳的新冠疫情防控是高效、有力的。2022 年一季度新冠病毒感染复阳率有显著下降，可能与深圳采取严格隔离措施，患者“已达到出院标准，继续予以留院或居家监测医学观察”的处理方法有关。本研究结果提示，临床医师应该关注新冠病毒感染者愈后安全，特别是少年在新冠病毒感染愈后 1 周内的易感期，需要注重提升该类人群愈后的安全防范意识，避免再次复阳。此外，本研究也表明，定期采用数据挖掘手段对新冠病毒感染者诊疗数据追踪分析，可发现不同时期、不同新冠病毒毒株感染者病情演变规律，为新冠病毒疫情防控提供科学的决策依据。

〔参考文献〕

〔1〕 殷紫薇，朱虹，涂乾．474 例新冠住院患者临床特征及重症危险因素分析〔J〕．江汉大学学报 (自然科学版)，2021，49(2)：5-11．

〔2〕 丁敬美，韩磊，王琳，等．200 例新冠肺炎住院患者转归影响因素分析〔J〕．解放军医院管理杂志，2020，27(6)：511-515．

〔3〕 滕辉，何兰，宋运娜．生物医学数据挖掘方法研究〔J〕．中国继续医学教育，2017，9(32)：22-24．

〔4〕 廖康生，卢洪洲．新型冠状病毒奥密克戎变异株的研究进展：对其科学防控措施的启示〔J〕．新发传染病电子杂志，2022，7(1)：1-5．

〔5〕 王彩红，姚晓文，王蓉，等．新冠病毒“奥密克戎亚变体 BA.5”的最新研究进展〔J〕．海南医学院学报，2022，28(20)：1521-1525．

〔6〕 吕莹，袁伟，施冬玲，等．2019 新型冠状病毒奥密克戎变异株感染者的临床特征分析〔J〕．中华传染病杂志，2022，40(5)：257-263．

〔7〕 张佳琦，刘国华，黄建安．新冠病毒奥密克戎变异株的特点与防控措施〔J〕．中国感染控制杂志，2022，21(8)：816-822．

〔8〕 陈芳．坚持“动态清零”不放松〔N〕．人民日报，2022-03-31(8)．

〔文章编号〕 1007-0893(2022)22-0004-05 DOI：10.16458/j.cnki.1007-0893.2022.22.002

β– 榄香烯介导 HIF–1α 对人食管 ESD 术后成纤维细胞的抑制作用

彭敦煌[1] 吴敬炬[2] 房太勇[1] 洪才发[1※]

（1. 福建医科大学附属第二医院，福建 泉州 362000；2. 福建医科大学第二临床医学院，福建 泉州 362000）

〔摘 要〕 **目的：**研究 β– 榄香烯通过介导缺氧诱导因子 –1α（HIF–1α）对人行食管内镜黏膜下剥离术（ESD）后成纤维细胞的抑制作用。**方法：**将 ESD 术后食管肉芽组织中提取、培养的成纤维细胞分为对照组（常规培养）、不同浓度 β– 榄香烯观察组（160 μmol • L^{-1}、320 μmol • L^{-1}、480 μmol • L^{-1}、640 μmol • L^{-1} 的 β– 榄香烯处理的 ESD 成纤维细胞），给药 0、24 h、48 h、72 h 后，用细胞增殖毒性检测法检测成纤维细胞生长抑制率，用流式细胞术检测成纤维细胞凋亡率，用蛋白质印迹法检测对照组和 480 μmol • L^{-1} β– 榄香烯观察组 HIF–1α 的表达水平。**结果：**β– 榄香烯浓度增加可减缓成纤维细胞增殖，成纤维细胞的凋亡率随着 β– 榄香烯浓度的增加而增加，HIF–1α 表达水平随着 β– 榄香烯浓度的增加而降低；用 HIF–1α 过表达质粒转染的成纤维细胞中的 HIF–1α 水平显著升高，通过 480 μmol • L^{-1} β– 榄香烯处理可逆转升高。**结论：**β– 榄香烯通过介导 HIF–1α 水平，减弱人食管成纤维细胞的增殖和加速成纤维细胞凋亡来抑制食管纤维化。

〔关键词〕 β– 榄香烯；食管狭窄；内镜黏膜下剥离术；成纤维细胞；缺氧诱导因子 –1α

〔中图分类号〕 R 571 〔文献标识码〕 A

〔收稿日期〕 2022 - 09 - 15

〔基金项目〕 福建省自然科学基金项目（2020J01242）

〔作者简介〕 彭敦煌，男，住院医师，主要从事食管早癌的防治研究。

〔※ 通信作者〕 洪才发（E-mail：allen1121@126.com；Tel：15980077626）