DISCOVID: discovering patterns of COVID-19 infection from recovered patients: a case study in Saudi Arabia

Tarik Alafif[1] · Alaa Etaiwi[2] · Yousef Hawsawi[3] · Abdulmajeed Alrefaei[4] ·
Ayman Albassam[1] · Hassan Althobaiti[1]

Abstract A respiratory syndrome COVID-19 pandemic has become a serious global concern. Still, a large number of people have been daily infected worldwide. Discovering COVID-19 infection patterns is significant for health providers towards understanding the infection factors. Current COVID-19 research works have not been attempted to discover the infection patterns, yet. In this paper, we employ an Association Rules Apriori (ARA) algorithm to discover the infection patterns from COVID-19 recovered patients' data. A non-clinical COVID-19 dataset is introduced and analyzed. A sample of recovered patients' data is manually collected in Saudi Arabia. Our manual computation and experimental results show strong associative rules with high confidence scores among males, weight above 70 kilograms, height above 160 centimeters, and fever patterns. These patterns are the strongest infection patterns discovered from COVID-19 recovered patients' data.

Tarik Alafif and Alaa Etaiwi contributed equally to this work.

✉ Tarik Alafif
  tkafif@uqu.edu.sa

  Alaa Etaiwi
  aketaiwi@gmail.com

  Yousef Hawsawi
  hyousef@kfshrc.edu.sa

  Abdulmajeed Alrefaei
  afrefaei@uqu.edu.sa

  Ayman Albassam
  asbassam@uqu.edu.sa

  Hassan Althobaiti
  hmthobaiti@uqu.edu.sa

[1] Computer Science Department, Jamoum University College, Umm Al-Qura University, Jamoum 25375, Makkah, Saudi Arabia

[2] Pathology and laboratory medicine Department, King Faisal Specialist Hospital and Research Center, Jeddah 21499, Makkah, Saudi Arabia

[3] Saudi Human Genome program-Jeddah Satellite Laboratory, Research Center, King Faisal Specialist Hospital and Research Center, Jeddah 21499, Makkah, Saudi Arabia

[4] Biology Department, Umm Al-Qura University, Jamoum 25375, Makkah, Saudi Arabia

Keywords COVID-19 · Infection patterns · Association rules apriori algorithm

1 Introduction

In December 2019, a new virus known as severe acute respiratory syndrome Coronavirus 2 (SARS-CoV-2) emerged initially in Wuhan city of Hubei province in China [1]. A few months later, in January 2020, the whole genome sequence of the SARS-CoV-2 virus was reported in China [1]. The lethal virus belongs to the Coronavirus family and therefore it is called COVID-19 [2, 3]. One of the particular properties of SARS-CoV-2 virus is spreading very rapidly, and therefore, the Worldwide Health Organization (WHO) announced the pandemic as a Public Health Emergency of International Concern (PHEIC)[2, 4]. According to Johns Hopkins statistics for SARS-CoV-2 virus, until March 20 in 2022, more than 468 million new cases were reported with over 6 million deaths worldwide.

On March 2020, the first confirmed case was reported in Saudi Arabia. On June 24 in 2021, Saudi Arabia was ranked as the fortieth affected country worldwide with over 749,471 confirmed cases and 8,485 SARS-CoV-2-related deaths.

Globally, SARS-Cov-2 causes no symptoms, mild or severe symptoms to patients through the virus replications. The patients are usually suffering from dry cough, fever, dyspnea, pneumonia, anosmia, fatigue, gastrointestinal manifestations, and ageusia [5, 6].

Data Mining (DM) techniques play a major role in many research fields. These techniques have the ability to solve complex tasks and extract useful knowledge from small and big data for many diseases including infectious diseases [7]. In addition, DM techniques can also assist medical practitioners in healthcare systems in the war on COVID-19 pandemic.

Our work is motivated by the success of the Association Rules Apriori (ARA) algorithm [8] in finding patterns in data. The ARA algorithm is a popular DM technique proposed by Agrawal in 1993 [9] where it scans a dataset to identify common itemsets and generate associative rules. It was commercially used for business transactions to identify interesting associations and patterns between customers and most frequently purchased items using minimum support and minimum confidence thresholds. The patterns are generated by association rules. Each association rule consists of a consequent and an antecedent to describe the patterns. A famous example of association rule mining is the market basket analysis which helps businesses to expand marketing strategies by providing awareness into which items are most frequently purchased by customers. In this research, we intend to employ the ARA algorithm to find the most frequent infection patterns from COVID-19 recovered patients' data.

Noticeably and luckily, there is a large number of infected people who have been recovered. Questions may be raised on how these people are infected? What are their health patterns? What are the most frequent infection patterns from those recovered patients? Mining the answers to these questions scientifically is important to find and understand the different infection patterns of COVID-19 patients. Therefore, understanding and finding the infection patterns of COVID-19 patients can assist medical practitioners in the behavior of the infection and spread public health awareness towards the COVID-19 pandemic. Meanwhile, discovering the infection patterns requires attaining infection and health information from recovered COVID-19 patients. However, existing public COVID-19 datasets lacks of infection and health information to help in discovering these patterns. Peiffer et al. [10] and Alafif et al. [11] support our statement by asserting that operations using machine learning and DM which they require a global collaboration and data sharing to tackle the COVID-19 pandemic problem. In addition to the lack of COVID-19 infectious information datasets, existing predictive COVID-19 methods such as in [12, 13], and [14] focus only on predicting the states of infected patients and

the spread of COVID-19 respectively rather than attempting to find their infection patterns. To address the limitations in the existing COVID-19 datasets and methods, in this research work, we contribute the following:

- We introduce our COVID-19 infection patterns (IP) dataset which includes a sample of recovered patients' data collected in Saudi Arabia.
- We apply the ARA algorithm to discover associative infection patterns from recovered COVID-19 patients' data using the IP dataset.
- We show manual computation results and verify them with experimental results.
- A discussion on the results and their relations with the results of existing works is provided.

The remainder of this paper is organized as follows. In Sect. 2, we describe and analyze the IP dataset. Then, we present our proposed method to discover infection patterns from COVID-19 patients in Sect. 3. Experimental implementation and results are provided in Sect. 4. Then, discussions on the results, limitations, and challenges are provided in Sect. 5. Finally, we conclude our work in Sect. 6.

2 IP dataset

This study is conducted on 131 Saudi and non-Saudi subjects using a survey method. The study includes COVID-19 positive cases who spent a quarantine period in an isolation facility. Negative cases, positive cases that were critical and subjected to hospital admission, none critical cases who's subjected to home isolation, and those with missing information were excluded in the survey. The survey questionnaire includes biographic data such as age, gender, nationality, smoking status, weight, height, blood type, and sport activity. It also includes a medical history such as multivitamins, medicine usage before infection, medicine usage after infection, symptoms after infection, chronic disease, previous infection of MERS-CoV, previous infection of COVID-19, and period of isolation. The IP dataset is publicly available on[1].

The collected data from the questionnaire are tabulated as shown in Table 1. The data is manually processed and converted to a binary format.

Our preliminary analysis for the collected data demonstrates that 21.4% of the study population's age is under 20 years old while the remaining 83.6% is above 20 years old (33.7% of them are between 30-40 years old while 25.5% are between age 20-30 years old). Although this indicates

---

[1] https://www.kaggle.com/tarikalafif/covid19-infection-patterns-dataset

	Pattern	Sub-pattern	No. of individual
Table 1 Questionnaires data	Age (years)	0–10	12
		11–20	16
		21–30	39
		31–40	37
		41–50	16
		51–60	8
		Above 61	3
	Gender	Male	82
		Female	49
	Smoking	Smoker	31
		Non-Smoker	100
	Body mass index	Underweight (Below 18.5)	14
		Normal (18.5–24.9)	33
		Overweight (25–29.9)	40
		Obese I (30–34.5)	21
		Obese II (35–39.9)	12
		Obese III (Above 40)	8
	Blood group	A +ve	32
		A -ve	3
		B +ve	28
		B -ve	3
		O +ve	57
		O-ve	1
		AB+ve	6
		AB-ve	3
	isolation period (days)	2	1
		3	1
		5	2
		9	12
		10	42
		11	41
		12	3
		13	3
		14	13
		15	2
		16	1
		18	1
		20	1
		21	2
		25	1
		32	1

that the incidents rate in adults is higher, the obesity factor possibly increases the incidents in younger populations [15]. Figure 1 shows the age distribution in our study population.

Among our data in the dataset, gender is studied. It is found that 64.1% of the total subjects are male (84 subjects) while 37.4% are female (49 subjects). This association is reported in China back in February 2020 [16]. Figure 2 shows the gender domination in the study population.

In the IP dataset, 43% (56 subjects) have O+ blood group while 24% (32 subjects) have A+. Our data show that subjects who have O+ groups are more exposed to have COVID-19 than other blood types. However, this
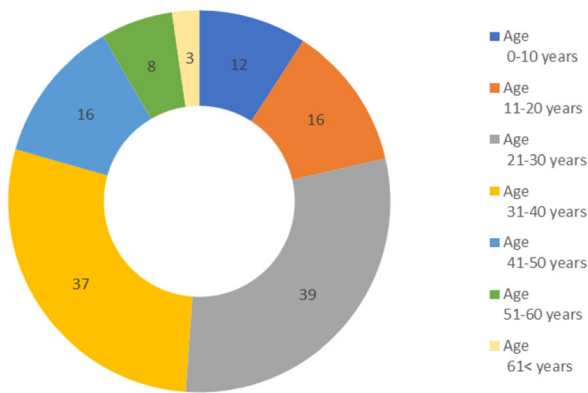
2828

Int. j. inf. tecnol. (October 2022) 14(6):2825–2838

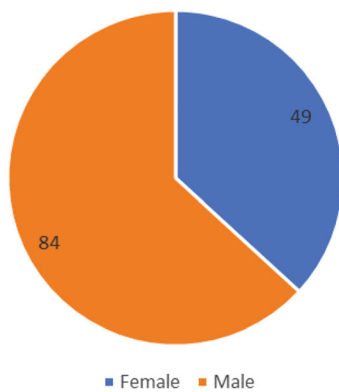**Fig. 1** Age distribution in the study population

Fig. 2 Gender domination in the study population

finding contradicts a similar study in Germany that excludes such a link [17]. Figure 3 shows the blood group in the study population.

Studies highlighted an insignificant relation between smoking and infection with COVID-19 [18]. As it is illustrated in our study, 23% of 131 subjects are smokers whereas is the remaining 77% are non-smokers. Figure 4 shows the percentage of smoker and non-smoker subjects in the study population.
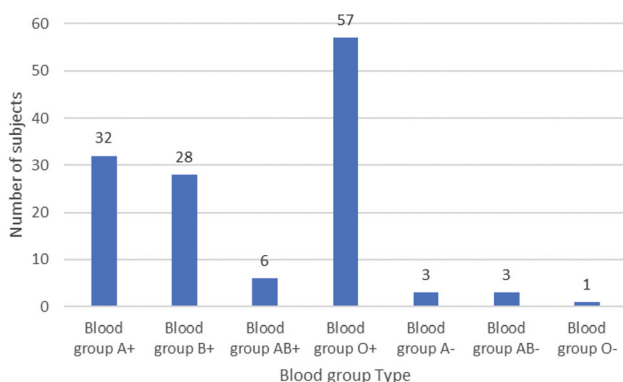

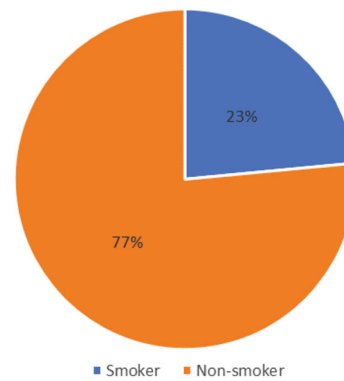**Fig. 3** Blood group type in the study population

Fig. 4 Smoker and non-smoker subjects in the study population

Moreover, most subjects develop symptoms like cough, fever, diarrhea, headache, fatigue, loss of taste and smell while few of them don't develop any symptoms as shown in Fig. 5. We notice that subjects' history with chronic diseases is not related to the infection pattern nor administration of multivitamins as supported in [19]. Figures 6 and 7 show the distribution of chronic disease and multivitamins administration in the study population respectively.

Due to COVID-19 precautionary measures, difficulties in accessing the electronic link, and elder subjects, both surveys and consents are conducted online using smartphones, tablets, and other electronic devices. A soft copy is sent to the subjects as a link. The actual subjects' size is 330 subjects. However, due to difficulties in obtaining their consent and conducting the survey, it has led to limiting the subjects' size to 131. These factors include but are limited to; wrong/unresponsive contacts' number, refusal to participate, subjects' unfamiliarity with their blood groups, medications, height, and weight. Furthermore, most of the isolation facilities are recently closed and COVID-19 positive cases were directed to home isolation. This prevents us from reaching a higher targeted sample size.
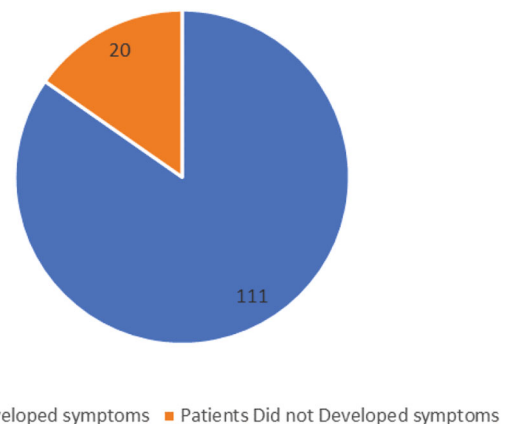

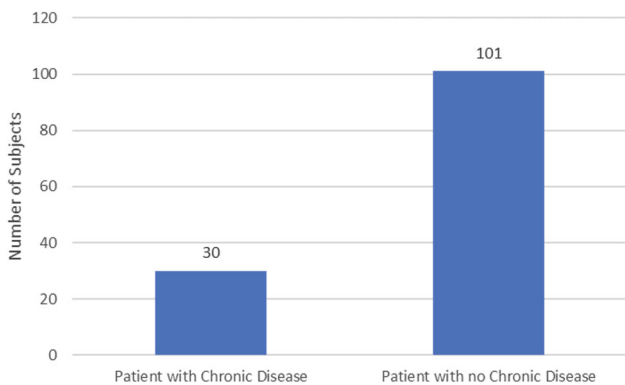**Fig. 5** Subjects with symptoms in the study population

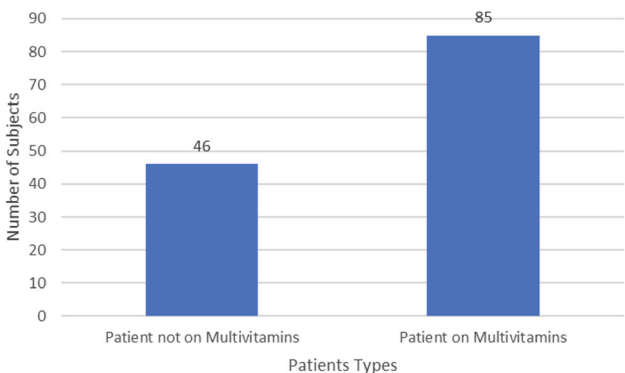**Fig. 6** Chronic disease in the study population

Fig. 7 Multivitamins administration in the study population

3 Methodology

In this section, we manually employ the ARA algorithm to extract and generate the strong associative rules and correlated patterns using the IP dataset. Figure 8 shows a pipeline of our proposed approach. The goal of using the ARA algorithm is to find the most frequent patterns relating to COVID-19 infection patterns.

Let COVID-19 infection patterns (P) where $P = \{p_1, p_2, \ldots, p_m\}$ is a set of binary patterns. Let D be the IP dataset. Each record of a infected patient (R) in D is represented as a binary vector, with $R[n] = 1$ if R has the pattern p_n. Otherwise, $R[n] = 0$ where n is the number of p. Let Q be a set of some patterns in P. We say that a record R satisfies X if for all pattern p_n in Q, $R[n] = 1$.

By applying the ARA algorithm, we imply $Q \longrightarrow p_n$, where Q is a set of some pattern in P, and p_n is an example of P that is not found in Q. The rule $Q \longrightarrow p_n$ is satisfied in the dataset D with a confidence threshold C where $0 \leq C \leq 1$ if and only if at least the confidence threshold of the records in D that satisfies Q tends to satisfy p_n.

Given the set of infection patterns p, we are interested in generating all the rules that satisfy minimum support *MinSup* and minimum confidence *MinConf* thresholds. The *MinSup* threshold helps to find the frequent patternsets in D that satisfy the union of patternsets in the consequent and antecedent of the rule. By using the *MinSup* threshold, we generate all combinations of patternsets and select only pattrnsets that have a higher value than the threshold. The *MinSup* threshold is computed as follows:

$$MinSup = count\ for\ patternset/Total\ number\ of\ R \qquad (1)$$

Then, we apply the *MinConf* threshold to find the frequent rules that occur among all the patternset in D. The *MinConf* threshold is computed as follows:

$$\text{MinConf} = \frac{\text{No. of support count in both } Q \text{ and } p_n}{\text{No. of support count in } Q} \qquad (2)$$

To generate strong associative rules, we first need to find all frequent patternsets. The first process is to join patterns to find L_k, a set of candidate k-patternsets C_k, where k is the number of elements in the frequent patternset. Then, a pruning process is performed by using the *MinSup* threshold to reduce the size of C_k in the k-frequent patternsets using the *MinSup* threshold. The value of *MinSup* threshold is selected to obtain higher frequent patternsets recursively. Our approach follows algorithm 1.

Fig. 8 The pipeline methodology of the proposed approach

2830

Int. j. inf. tecnol. (October 2022) 14(6):2825–2838

Algorithm 1 : Discovering patterns of COVID-19 infection from recovered patients data.

Input:

    COVID-19 infection patterns from recovered patients data.

  1: Let $k = 1$

  2: Generate frequent patternset of length 1.

  3: **for each** frequent patternset **do**

  4:    Generate length $(k + 1)$ candidate Patternsets from length k frequent patternsets.

  5:    Prune candidate patternsets which contain subsets of length k that are infrequent.

  6:    Count the support of each candidate by scanning the D.

  7:    Eliminate candidates are infrequent and leave the frequent patternsets.

  8: **end for**

  9: Compute the confidence scores to obtain only strong COVID-19 infection patternset associative rules.

Output:

    Generated strong COVID-19 infection patternsets associative rules.

Given the IP dataset in Table 2 which contains of patient identification (PID) and a list of patterns_IDs. Each patient may have a different list of Pattern_IDs. The pattern_IDs consists of letters in an alphabetical order which correspond to the infection patterns p respectively in the Excel sheet. The alphabetical letters are used to ease the manual computation and readability (Tables 2 and 3).

First, we scan the whole dataset D in Table 2 for counting the support of each candidate in the Pattern_IDs in the first pass. We call the first candidates C_1 or 1-freqnt patternset as shown in Table 4. After finding the support count for C_1 candidates, the support count are compared with the *MinSup* threshold. We set the *MinSup* threshold to 40% to filter out weak candidates that do not satisfy the *MinSup* threshold. We call these candidates L_1 as shown in Table 5. Similarly, the same procedure is followed to compute 2-frequent patternsets (C_2, L_2), 3-frequent patternsets (C_3,L_3), and 4-frequent patternsets (C_4,L_4) as shown in Tables 3, 6, 7, 8, 9, 10, and 11. The ARA algorithm stops to join and generate more candidates in the fifth pass since the 4-frequent patternsets do not satisfy the -*MinSup* threshold.

After pruning the infrequent patternsets, we reach the 3-frequent patternsets L_3 that satisfy the *MinSup* threshold. Then, we generate the association rules by generating all non-empty subsets s of frequent patternsets from L_3 and output the rules $s \longrightarrow (l - s)$. From Table 9, we see that L_3 consists of $l_1 = \{I, L, M\}$ and $l_2 = \{I, M, DA\}$. Then, we

generate all possible rules and use Equation 2 to compute the confidence for each rule as shown in Table 12.

We set the *MinConf* threshold to 60% to exclude weak generated rules from Table 12. Then, the strong generated rules that satisfy the *MinConf* threshold are as follows:

- $\{I\} \longrightarrow \{L, M\}$
- $\{L\} \longrightarrow \{I, M\}$
- $\{M\} \longrightarrow \{I, L\}$
- $\{I, L\} \longrightarrow \{M\}$
- $\{I, M\} \longrightarrow \{L\}$
- $\{L, M\} \longrightarrow \{I\}$
- $\{I, M\} \longrightarrow \{DA\}$
- $\{I, DA\} \longrightarrow \{M\}$
- $\{M, DA\} \longrightarrow \{I\}$

We substitute the alphabetical letters with the infection patterns. Then, we obtain the following rules:

- $\{Male\} \longrightarrow \{(Weight > 70), (Height > 160)\}$
- $\{(Weight > 70)\} \longrightarrow \{Male, (Height > 160)\}$
- $\{(Height > 160)\} \longrightarrow \{Male, (Weight > 70)\}$
- $\{Male, (Weight > 70)\} \longrightarrow \{(Height > 160)\}$
- $\{Male, (Height > 160)\} \longrightarrow \{(Weight > 70)\}$
- $\{(Weight > 70), (Height > 160)\} \longrightarrow \{Male\}$
- $\{Male, (Height > 160)\} \longrightarrow \{Fever\}$
- $\{Male, Fever\} \longrightarrow \{(Height > 160)\}$
- $\{(Height > 160), Fever\} \longrightarrow \{Male\}$

As shown from the above generated associative rules, we obtain eight-strong generated rules. We discover four

Table 2 List of pattern IDs for each patient's record in the dataset

ID	List of pattern IDs	ID	List of pattern IDs
1	E, I, L, M, Q, V, Z, AO, AQ, AW, BV, CW, CY, CZ, DA, DB, DC, DS	67	D, I, L, M, S, T, Z, AD, BV, CA, CB, DF, DK, DS
2	C, I, K, M, O, V, Z, AE, AX, BS, CC, CL, DA, DD, DE, DF, DS	68	D, J, O, T, DA, DB, DT
3	D, J, K, L, R, W, CY, DC, DE, DF, DG, DJ, DS	69	D, I, L, M, S, CY, DA, DB, DS
4	B, J, K, N, W, DA, DT	70	C, I, M, O, T, Z, DA, DS
5	E, I, K, M, O, W, Z, AV, BV, DA, DB, DE, DF, DH, DI, DT	71	C, I, L, M, P, T, Z, DE, DF, DS
6	E, I, K, L, M, O, Z, DC, DT	72	H, I, L, M, Q, T, AC, AD, AO, AU, BH, BK, BQ, BR, DT
7	E, I, K, L, M, R, CY, DT	73	O, I, L, M, P, T, BV, CI, CM, CY, DA, DB, DD, DI, DJ, DK, DT
8	G, I, K, L, M, P, Z, AD, BT, BU, CI, CY, DT	74	E, I, M, O, T, AV, DA, DD, DM, DS
9	F, J, K, O, AM, CC, CD, CY, DA, DC, DL, DT	75	B, I, K, S, V, Y, DB, DS
10	F, I, K, L, M, Q, Z, AF, BO, BV, CE, CZ, DT	76	E, I, K, L, M, P, V, DB, DS
11	F, I, K, L, M, R, Z, AC, AV, BP, BQ, BR, CI, CK, CL, CY, DA, DC, DK, DT	77	E, I, K, L, M, P, V, Z, BV, CM, CY, DA, DM, DT
12	B, J, K, O, CY, DA, DC, DT	78	E, I, K, L, M, P, V, Z, BV, CM, DA, DC, DT
13	B, J, K, P, CX, DI, DT	79	G, J, K, L, Q, V, AD, AN, AQ, AU, BM, BP, BQ, BR, CL, CQ, DA, DB, DE, DF, DI, DM, DT
14	D, I, K, L, M, P, BV, CI, CY, DA, DB, DC, DT	80	D, J, K, N, V, AS, BV, CI, CY, DA, DT
15	C, I, K, O, Z, AC, AP, AQ, CJ, DA, DI, DT	81	D, I, K, L, M, Q, V, Z, CY, DA, DD, DS
16	B, J, K, O, AP, AQ, BV, CI, DA, DI, DT	82	B, I, K, S, V, DA, DS
17	D, I, K, M, O, AP, AQ, DT	83	D, I, K, L, M, P, V, Z, DS
18	D, J, K, O, Z, AP, AQ, BV, DB, DK, DS	84	C, I, K, M, O, V, Z, CY, DD, DS
19	D, J, K, P, AL, AP, AQ, DS	85	F, I, K, L, M, S, V, AJ, BC, BD, BE, CF, CM, CY, CZ, DA, DR, DT
20	E, I, K, L, M, P, Z, AV, BV, BZ, CB, DA, DB, DC, DD, DE, DF, DS	86	E, I, K, L, M, P, V, Z, CM, DA, DJ, DK, DL, DT
21	E, I, K, L, M, P, Z, AV, BV, CB, CH, CY, DA, DC, DD, DE, DF, DM, DS	87	D, I, K, M, P, V, Z, DT
22	D, I, K, M, O, Z, DE, DF, FK, DS	88	E, J, K, M, N, V, AI, CV, CY, DB, DC, DE, DF, DS
23	C, J, K, O, Z, BV, CB, DA, DS	89	C, I, K, P, V, AP, AQ, CY, DI, DT
24	C, J, K, O, Z, BV, CB, DA, DS	90	B, I, K, N, V, AP, AQ, DA, DI, DT
25	B, I, K, N, AV, CB, DA, DB, DS	91	G, J, K, L, Q, V, AD, BQ, CB, CI, CN, DA, DT
26	B, I, K, N, Z, AV, CD, DA, DB, DN, DT	92	F, I, K, L, M, Q, V, BV, DA, DS
27	D, J, K, M, O, AN, AV, BM, CB, DA, DT	93	C, I, K, L, R, V, Z, DA, DE, DF, DT
28	D, I, K, L, R, Z, CK, DA, DT	94	D, I, K, L, M, P, V, BV, CY, DA, DB, DD, DS
29	D, J, K, L, Q, Z, BV, DS	95	F, I, K, L, M, Q, V, AC, BK, DA, DB, DF, DS
30	F, J, K, L, Q, BV, DB, DS	96	E, I, K, L, M, P, V, Z, BV, DB, DN, DT
31	E, J, K, Q, Z, DM, DS	97	F, I, K, L, M, P, V, BG, DA, DC, DT
32	D, J, K, O, AP, BX, CZ, DA, DB, DC, DD, DM, DS	98	E, I, K, L, M, R, V, Z, AG, BA, BB, BR, CI, CY, DC, M, DS
33	E, I, K, L, M, Q, BY, CZ, DC, DD	99	D, I, K, L, M, P, V, Z, CI, CY, DF, DS
34	D, I, L, M, P, Z, BV, CY, DA, DK, DT	100	D, J, K, P, V, AN, BM, CY, DA, DB, DC, DT
35	E, J, L, Q, AP, AQ, AW, CI, DA, DB, DC, DD, DI, DT	101	B, J, K, N, V, DS
36	E, J, L, Q, Z, BN, BV, DA, DB, DF, DM, DS	102	D, J, K, L, V, CY, DA, DB, DC, DT
37	D, I, M, O, DT	103	B, I, K, O, V, AE, AZ, DT
38	F, J, K, L, M, Q, U, AV, BV, CY, DT	104	E, I, K, L, M, P, V, AW, BV, DA, DE, DF, DK, DS
39	E, J, K, L, M, P, U, BV, DB, DC, DS	105	E, I, K, L, S, V, BV, BW, DA, DB, DH, DT
40	E, I, K, L, M, P, U, CL, DA, DC, DD, DE, DI, DS	106	E, I, K, L, M, P, V, AV, CB, DA, DT
41	E, J, L, M, Q, U, Z, AQ, AS, AU, BF, GY, DA, DB, DM, DT	107	D, I, K, L, M, Q, V, AL, AQ, AY, CI, CY, DE, DF, DK, DS
42	G, I, L, M, R, U, Z, DS	108	H, J, K, L, S, V, AC, AD, AO, AQ, AT, AV, CS, CT, DA, DD, DS

2832

Int. j. inf. tecnol. (October 2022) 14(6):2825–2838

Table 2 continued

ID	List of pattern IDs	ID	List of pattern IDs
43	E, I, M, O, U, DS	109	D, I, K, L, M, R, V, CI, CL, CY, DA, DB, DD, DS
44	E, I, K, L, R, X, Z, CL, CM, DT	110	E, J, K, P, V, AP, CL, CU, CY, CZ, DA, DB, DC
45	C, I, K, M, N, X, BV, DA, DC, DS	111	D, J, K, O, V, Z, DA, DB, DC, DM, DS
46	E, I, L, M, P, X, AW, BV, CY, DA, DB, DE, DF, DT	112	F, J, K, P, V, AV, AW, CY, CZ, DA, DB, DC, DS
47	D, J, K, O, T, Z, AP, AT, AU, BV, CM, DA, DC, DT	113	D, I, M, O, V, Z, AW, BV, DA, DB, DC, DM, DT
48	C, J, K, O, T, Z, BG, BL, CY, DA, DC, DT	114	D, I, M, O, V, Z, AQ, DA, DC, DK, DT
49	G, J, K, L, M, Q, T, BK, BS, CL, DA, DC, DP, DT	115	D, I, M, O, V, DA, DT
50	E, I, K, L, M, Q, T, DA, DE, DF, DK, DM, DT	116	D, I, M, P, V, BV, CB, CL, DA, DB, DS
51	D, I, K, L, M, Q, T, AC, DT	117	D, I, M, N, V, Z, BV, DA, DS
52	F, J, K, L, S, T, AD, AK, AN, AT, BJ, DS	118	E, J, P, V, AV, BV, CM, CY, DA, DD, DK, DT
53	E, J, K, P, T, AH, AU, BI, BV, CY, DI, DS	119	E, I, L, M, P, V, DA, DE, DF, DM, DT
54	D, J, K, M, O, T, BV, CB, CL, DA, DS	120	D, I, L, M, O, V, DA, DS
55	G, I, K, N, L, M, Q, T, AC, AD, AT, BK, BQ, CB, CI, CN, DA, DC, DF, DT	121	D, I, L, M, O, V, Z, BV, CI, DA, DD, DT
56	F, I, K, L, M, Q, T, Z, AC, AD, BK, BQ, DD, DS	122	F, J, L, Q, V, Z, CY, DA, DE, DF, DT
57	D, J, K, L, M, P, T, Z, AH, AT, AU, BV, BW, DA, DE, DF, DH, DI, DS	123	H, J, M, O, V, CL, DA, DS
58	D, I, K, L, M, P, T, Z, AX, CX, DN, DS	124	E, I, L, M, N, V, AP, AQ, BV, DC, DS
59	E, J, K, L, R, T, CD, CM, DA, DS	125	C, J, K, N, V, BV, DS
60	F, I, K, L, M, P, T, Z, AG, AO, AT, BK, BR, BV, DA, DB, DP, DQ, DT	126	C, I, K, N, V, BV, DT
61	E, J, K, P, T, AP, AR, BV, CO, CY, CZ, DA, DC, DD, DM	127	B, I, K, N, V, BV, DT
62	F, J, K, L, R, T, AC, AD, AV, CP, CR, CY, CZ, DA, DB, DC	128	C, I, K, N, V, BV, DT
63	E, I, K, M, O, T, AV, BV, CY, CZ, DA, DB, DC, DT	129	C, I, K, O, V, BV, DT
64	F, J, K, L, R, T, AP, AR, BV, CM, CY, DB, DC, DD, DM	130	C, I, K, L, M, P, Z, BV, CY, DA, DC, DT
65	G, I, M, P, T, Z, AV, AW, DT	131	E, J, K, P, BV, DA, DF, DM, DS
66	G, J, P, T, Z, AS, AU, BR, DK, DT	-	-

strong COVID-19 infection patterns from COVID-19 recovered patients. The infection patterns are *Male*, $(Weight > 70)$, $(Height > 160)$, and *Fever*. These are the most frequent and strongest patterns obtained by the ARA algorithm.

From the above strong generated associative rules, one can notice that the first six rules have the same patterns *Male*, $(Weight > 70)$, $and (Height > 160)$. The last three rules contain *Fever* pattern in addition to *Male* and $(Height > 160)$ patterns.

From the rule $\{Male\} \longrightarrow \{(Weight > 70), (Height > 160)\}$, if a recovered patient is a male, he tends to have a pattern of a weight more than 70 kilogram and a pattern of height more than 160 centimeters. Similarly, the rules $\{(Weight > 70)\} \longrightarrow \{Male, (Height > 160)\}$, $\{(Height > 160)\} \longrightarrow \{Male, (Weight > 70)\}$, $\{Male, (Weight > 70)\} \longrightarrow \{(Height > 160)\}$, $\{Male, (Height > 160)\} \longrightarrow \{(Weight > 70)\}$, and $\{(Weight > 70), (Height > 160)\} \longrightarrow \{Male\}$ provide the same strong patterns that are strongly associated to each other.

For the rule $\{Male, (Height > 160)\} \longrightarrow \{Fever\}$, if a recovered patient is a male and his height is more than 160 centimeters, he tends to have a pattern of fever. Similarly, the rules $\{Male, Fever\} \longrightarrow \{(Height > 160)\}$ and $\{(Height > 160), Fever\} \longrightarrow \{Male\}$ provide the same strong patterns that are strongly associated to each other.

4 Experiment

The ARA algorithm experiment is conducted using Google Co-laboratory with Apriori Python library on a Windows 10 Pro with 64-bit 8-core and Intel(R) Core(TM) i7-7700K CPU @ 4.20GHz with 32.0 GB of RAM.

The goal of the experiment is to verify the results obtained from our manual computation. The *MinSup* and the *MinConf* thresholds are used in the experiment. The same results of the generated strong associative rules are obtained in addition to the non-smoking pattern. The experiment runs in 1.1 seconds.

Table 3 Scanning the datasest for counting each 2-frequent patternset (C2)

Patternset	Support Count	Patternset	Support Count	Patternset	Support Count
{BV, CY}	16	{J, CY}	17	{M, V}	16
{BV, DA}	32	{J, DA}	33	{M, Z}	33
{BV, DS}	17	{J, DS}	18	{M, BV}	28
{BV, DT}	16	{J, DT}	12	{M, CY}	24
{CY, DA}	27	{K, I}	62	{M, DA}	44
{CY, DS}	10	{K, J}	41	{M, DS}	28
{CY, DT}	8	{K, L}	53	{M, DT}	22
{DA, DS}	23	{K, M}	49	{P, V}	19
{DS, DT}	0	{K, P}	29	{P, Z}	17
{I, J}	0	{K, V}	44	{P, BV}	19
{I, L}	52	{K, Z}	36	{P, CY}	12
{I, M}	64	{K, BV}	35	{P, DA}	22
{I, P}	28	{K, CY}	27	{P, DS}	15
{I, V}	42	{K, DA}	54	{P, DT}	15
{I, Z}	38	{K, DS}	42	{V, Z}	15
{I, BV}	30	{K, DT}	50	{V, BV}	17
{I, CY}	22	{L, M}	52	{V, CY}	18
{I, DA}	48	{L, P}	26	{V, DA}	38
{I, DS}	24	{L, V}	30	{V, DS}	19
{I, DT}	28	{L, Z}	32	{V, DT}	13
{J, L}	19	{L, BV}	27	{Z, BV}	20
{J, M}	10	{L, CY}	25	{Z, CY}	13
{J, P}	12	{L, DA}	43	{Z, DA}	27
{J, V}	16	{L, DS}	25	{Z, DS}	19
{J, Z}	13	{L, DT}	22	{Z, DT}	19
{J, BV}	18	{M, P}	28	{-}	-

5 Discussion

The ARA algorithm is robust in finding and extracting patterns from data with the use of minimum support and minimum confidence thresholds. The ARA algorithm has shown strong associative COVID-19 infection patterns from the manual and the experimental computations using the IP dataset. As a result, male, weight above 70 kilograms, height above 160 centimeters, and fever are the strongest COVID-19 recovery patterns.

To support and associate our findings with the medical findings from the literature, some medical research works have associated some patterns for COVID-19 infection. The severity of symptoms in COVID-19 active case and mortality tendency is two-fold and a half higher than females in China [20]. In Italy, 70% of COVID-19 mortality is seen in men while the death rate in Wuhan is 1.1% higher in males than females [21, 22]. Although the ministry of health reported on the 20th of April that 55% of the reported active cases are females, we believe that this is related to public compliance to precautionary measures rather than a significant link to COVID-19. Similar findings is recorded in Oman on May 2020, where 70% of active cases are men as well [23]. Our generated associative rules have detected a pattern linking height and weight to COVID-19 infection. In 2019, the CDC reported that obesity increases the risk of severe illness for COVID-19 and overweight triple the risk of hospitalization [24]. Kompaniyets et al. [25] reported that risks of hospitalization, intensive care unit admission, invasive mechanical ventilation, and death are increased with increasing Body Mass Index (BMI). Our finding in the current study is consistent with the previous studies. We find that overweight individuals have an increased risk of having the COVID-19 infection. Other studies linked BMI above 25 kg/m2 as a strong disposing factor for COVID-19 infection [26].

Our results demonstrate a link between the infection of COVID-19 and non-smokers in our population. Although the same was observed in the UK, they found that the mortality rate is high in the infected smokers than the non-smokers [27]. Furthermore, Smoking is considered a risk factor for the progression of COVID-19 symptoms [28].

Table 4 Scanning the datasest for counting each 1-frequent patternset (C1)

Sub-pattern IDs	Support count	Sub-pattern IDs	Support count	Sub-pattern IDs	Support count
{B}	12	{AP}	14	{CF}	1
{C}	16	{AQ}	15	{CG}	0
{D}	39	{AR}	2	{CH}	1
{E}	37	{AS}	3	{CI}	14
{F}	16	{AT}	6	{CJ}	1
{G}	8	{AU}	7	{CK}	2
{H}	3	{AV}	16	{CL}	11
{I}	82	{AW}	8	{CM}	10
{J}	49	{AX}	2	{CN}	21
{K}	100	{AY}	1	{CO}	1
{L}	72	{AZ}	1	{CP}	1
{M}	75	{BA}	1	{CQ}	1
{N}	14	{BB}	1	{CR}	1
{O}	33	{BC}	1	{CS}	1
{P}	40	{BD}	1	{CT}	1
{Q}	23	{BE}	1	{CU}	1
{R}	12	{BF}	1	{CV}	1
{S}	8	{BG}	2	{CW}	1
{T}	28	{BH}	1	{CX}	2
{U}	6	{BI}	1	{CY}	1
{V}	57	{BJ}	1	{CZ}	1
{W}	3	{BK}	6	{DA}	81
{X}	3	{BL}	1	{DB}	38
{Y}	1	{BM}	3	{DC}	37
{Z}	51	{BN}	1	{DD}	19
{AA}	0	{BO}	1	{DE}	18
{AB}	0	{BP}	2	{DF}	23
{AC}	10	{BQ}	6	{DG}	1
{AD}	10	{BR}	6	{DH}	3
{AE}	2	{BS}	2	{DI}	12
{AF}	1	{BT}	1	{DJ}	3
{AG}	1	{BU}	1	{DK}	13
{AH}	2	{BV}	49	{DL}	2
{AI}	1	{BW}	2	{DM}	17
{AJ}	1	{BX}	1	{DN}	3
{AK}	1	{BY}	1	{DO}	0
{AL}	2	{BZ}	1	{DP}	2
{AM}	2	{CA}	1	{DQ}	1
{AN}	4	{CB}	13	{DR}	1
{AO}	4	{CC}	2	{DS}	59
{AP}	14	{CD}	2	{DT}	68
{AQ}	15	{CE}	1	{-}	-

In terms of the height pattern, our study shows that tall individuals are more likely to have COVID-19. This finding is in line with other reported studies by Desmet et al. which finds that individuals with height above 182 cm (6ft) tall most probably have doubles chances of being diagnosed with COVID-19 [29].

Knowing the first symptom of COVID-19 infection plays such an important role in the prevention of virus spread and patient's treatment plan. Several clinical studies have shown that almost all infected patients reported fever as the first symptom of the COVID-19 diseases [30, 31]. For example, in China, 98% of infected individuals have

Table 5 Comparing the support count of 1-frequent patternset (L1) with the MinSup count

L1 Sub-pattern IDs	Support count
{I}	82
{J}	49
{L}	72
{K}	100
{M}	75
{P}	40
{V}	57
{Z}	51
{PV}	49
{CY}	40
{DA}	81
{DS}	43
{DT}	41

Table 6 Comparing the support count of 2-frequent patternset (L2) with the MinSup count

Pattern set	Support count
{ I, L }	52
{ I, M }	64
{ I, V }	42
{ I, DA }	48
{ K, I }	62
{ K, J }	41
{ K, L }	53
{ K, M }	49
{ K, V }	44
{ K, DA }	54
{ K, DS }	42
{ K, DT }	50
{ L, M }	52
{ L, DA }	43
{ M, DA }	44
{ I, DT }	46
{ DA, DT }	47
{ BV, CY }	16
{ BV, DA }	32
{ BV, DS }	25
{ BV, DT }	23
{ CY, DA }	27
{ CY, DS }	14
{ CY, DT }	23
{ DA, DS }	32
{ DA, DT }	47

Table 7 Generating 3-frequent patternset C3 from L2

Pattern set
{I, L, M}
{I, M, V}
{I, DA, L}
{I, M, DA}
{L, M, DA}
{I, L, V}
{ I, V, DT }
{ I, DA, DT }
{ L, DA, DT }
{ I, L, DT }
{ I, M, DT }
{ K, I, J }
{ K, I, M }
{ K, I, V }
{ K, I, DA }
{ K, I, DS }
{ K, I, DT }
{ K, J, L }
{ K, J, M }
{ K, L, M }
{ K, L, V }
{ K, L, DA }
{ K, L, DS }
{ K, L, DT }
{ K, M, V }
{ K, M, DA }
{ K, M, DS }
{ K, M, DT }
{ K, V, DA }
{ K, V, DS }
{ K, V, DT }
{ K, DA, DS }
{ K, DA, DT }
{ K, DS, DT }
{ M, DA, DT }

suffered from fever, and in Italy, more than 72% reported the same symptom [32]. Furthermore, the Centre of Disease Control and Prevention in the US lists fever or chills as the first symptoms appear in 2-14 days after infection by the virus (https://www.cdc.gov/coronavirus/2019-ncov/symptoms-testing/symptoms.html). Also, fever lasts longer in the infected individual who is admitted to intensive care

2836

Int. j. inf. tecnol. (October 2022) 14(6):2825–2838

Table 8 Scanning the datasest for counting each 3-frequent patternset C3

Patternset	Support count
{I, L, M}	50
{I, M, V}	31
{ I, DA, L}	30
{I, M, DA}	40
{L, M, DA}	33
{I, L, V}	24
{I, M, DT }	32
{ I, V, DT }	22
{ I, L, DT }	28
{ I, DA, DT }	27
{K, I, DA }	33
{K, I, DS }	22
{K, I, DT }	33
{K, J, L }	13
{K, J, M }	7
{K, L, M }	38
{K, L, V }	24
{K, L, DA }	32
{K, L, DS }	24
{K, L, DT }	27
{K, M, V }	20
{K, M, DA }	29
{ K, M, DS }	23
{K, M, DT }	24
{K, V, DA }	23
{K, V, DS }	21
{K, V, DT}	22
{K, DA, DS}	22
{K, DA, DT}	33
{K, V, DT}	22
{K, DS, DT}	0
{ L, DA, DT }	27
{ M,DA, DT }	25

Table 9 Comparing the support count of 3-frequent patternset (L3) with the MinSup count

Patternset
{ I, L, M }
{ I, M, DA }

Table 10 Generating 4-frequent patternset C4 from L3

Patternset
{ I, L, M, DA }

Table 11 Scanning the datasest for counting each 4-frequent patternset C4

Patternset	Support Count
{ I, L, M, DA }	29

Table 12 Generated rules from 3-frequent patternsets L_3

Generated rule	Confidence
$\{I\} \longrightarrow \{L, M\}$	50 / 82 = 0.6097 = 60.97%
$\{L\} \longrightarrow \{I, M\}$	50 / 72 = 0.6944 = 69.44%
$\{M\} \longrightarrow \{I, L\}$	50 / 75 = 0.6666 = 66.66%
$\{I, L\} \longrightarrow \{M\}$	50 / 52 = 0.9615 = 96.15%
$\{I, M\} \longrightarrow \{L\}$	50 / 64 = 0.7812 = 78.12%
$\{L, M\} \longrightarrow \{I\}$	50 / 52 = 0.9615 = 96.15%
$\{I\} \longrightarrow \{M, DA\}$	40 / 82 = 0.4878 = 48.78%
$\{M\} \longrightarrow \{I, DA\}$	40 / 75 = 0.5333 = 53.33%
$\{DA\} \longrightarrow \{I, M\}$	40 / 81 = 0.4938 = 49.38%
$\{I, M\} \longrightarrow \{DA\}$	40 / 64 = 0.625 = 62.5%
$\{I, DA\} \longrightarrow \{M\}$	40 / 48 = 0.8333 = 83.33%
$\{M, DA\} \longrightarrow \{I\}$	40 / 44 = 0.9090 = 90.90%

units after the start of the COVID-19 symptoms [33]. Clearly, fever should be considered as the first sign/indicator of the onset of COVID-19 infection.

6 Conclusion

Discovering infection patterns is significant for health providers towards understanding the infection factors. In this research work, we propose to use the ARA algorithm to discover infection patterns from recovered COVID-19 patients' data. A non-clinical COVID-19 dataset is introduced and analyzed, which consists of COVID-19 Infection patterns. A sample of recovered patients' data is manually collected in Saudi Arabia. Our manual computation and experimental results show strong associative rules with high confidence scores among males, weight above 70 kilograms, height above 160 centimeters, and fever patterns. These patterns are the strongest infection patterns discovered from COVID-19 recovered patients'

data. In future, we will work on the discovery of COVID-19 recovery and death patterns.

Availability of data and materials Applicable.

Code availability Not applicable.

Declarations

Conflict of interest Not applicable.

Ethical approval Applicable.

Consent to participate Applicable.

Consent for publication Applicable.

References

1. World Health Organization and others: Coronavirus disease 2019 (covid- 19): situation report, 72 (2020)

2. Phelan AL, Katz R, Gostin LO (2020) The novel coronavirus originating in Wuhan, China: challenges for global health governance. JAMA 323(8):709–710

3. Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, Zhao X, Huang B, Shi W, Lu R et al (2020) A novel coronavirus from patients with pneumonia in China, 2019. N Engl J Med 382:727–733

4. Ebrahim SH, Memish ZA (2020) Covid-19: preparing for superspreader potential among Umrah pilgrims to Saudi Arabia. Lancet (London, England) 395(10227):48

5. Hoffman T, Nissen K, Krambrich J, Rönnberg B, Akaberi D, Esmaeilzadeh M, Salaneck E, Lindahl J, Lundkvist Å (2020) Evaluation of a COVID-19 IGM and IGG rapid test; an efficient tool for assessment of past exposure to SARS-COV-2. Infect Ecol Epidemiol 10(1):1754538

6. Long Q-X, Tang X-J, Shi Q-L, Li Q, Deng H-J, Yuan J, Hu J-L, Xu W, Zhang Y, Lv F-J et al (2020) Clinical and immunological assessment of asymptomatic SARS-COV-2 infections. Nat Med 26(8):1200–1204

7. Peiffer-Smadja N, Rawson TM, Ahmad R, Buchard A, Pantelis G, Lescure F-X, Birgand G, Holmes AH (2019) Machine learning for clinical decision support in infectious diseases: a narrative review of current applications. Clin Microbiol Infect 26:584–595

8. Agrawal R, Srikant R et al (1994) Fast algorithms for mining association rules. Proc Int Conf Very Large Data Bases 1215:487–499

9. Agrawal R, Imieliński T, Swami A (1993) Mining association rules between sets of items in large databases. In: Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, pp 207–216

10. Peiffer-Smadja N, Maatoug R, Lescure F-X, D'Ortenzio E, Pineau J, King J-R (2020) Machine learning for COVID-19 needs global collaboration and data-sharing. Nat Mach Intell 2:293–294

11. Alafif T, Tehame AM, Bajaba S, Barnawi A, Zia S (2021) Machine and deep learning towards COVID-19 diagnosis and treatment: survey, challenges, and future directions. Int J Environ Res Public Health 18(3):1117

12. Alafif T, Alotaibi R, Albassam A, Almudhayyani A (2021) On the prediction of isolation, release, and decease states for COVID-19 patients: a case study in South Korea. ISA Trans. https://doi.org/10.1016/j.isatra.2020.12.053

13. Iwendi C, Bashir AK, Peshkar A, Sujatha R, Chatterjee JM, Pasupuleti S, Mishra R, Pillai S, Jo O (2020) COVID-19 patient health prediction using boosted random forest algorithm. Front Public Health 8:357

14. Sujath R, Chatterjee JM, Hassanien AE (2020) A machine learning forecasting model for COVID-19 pandemic in India. Stoch Environ Res Risk Assess 34:959–972

15. Kass DA, Duggal P, Cingolani O (2020) Obesity could shift severe COVID-19 disease to younger ages. Lancet (London, England) 395:1544–1545

16. Zhang J-J, Dong X, Cao Y-Y, Yuan Y-D, Yang Y-B, Yan Y-Q, Akdis CA, Gao Y-D (2020) Clinical characteristics of 140 patients infected with SARS-COV-2 in Wuhan, China. Allergy 75:1730–1741

17. Latz CA, DeCarlo C, Boitano L, Png CM, Patell R, Conrad MF, Eagleton M, Dua A (2020) Blood type and outcomes in patients with COVID-19. Ann Hematol 99(9):2113–2118

18. Cai H (2020) Sex difference and smoking predisposition in patients with COVID-19. Lancet Respir Med 8(4):20

19. Singhal T (2020) A review of coronavirus disease-2019 (COVID-19). Indian J Pediatr 87:281–286

20. Jin J-M, Bai P, He W, Wu F, Liu X-F, Han D-M, Liu S, Yang J-K (2020) Gender differences in patients with COVID-19: focus on severity and mortality. Front Public Health 8:152

21. Remuzzi A, Remuzzi G (2020) COVID-19 and Italy: what next? Lancet 395(10231):1225–1228

22. Groban L, Wang H, Sun X, Ahmad S, Ferrario CM (2020) Is sex a determinant of COVID-19 infection? Truth or myth? Curr Hypertens Rep 22(9):1–12

23. Khamis F, Al Rashidi B, Al-Zakwani I, Al Wahaibi AH, Al Awaidy ST (2020) Epidemiology of COVID19 infection in Oman: analysis of the first 1304 cases. Oman Med J 35(3):141

24. Simonnet A, Chetboun M, Poissy J, Raverdy V, Noulette J, Duhamel A, Labreuche J, Mathieu D, Pattou F, Jourdain M et al (2020) High prevalence of obesity in severe acute respiratory coronavirus-2 (SARS-COV-2) requiring invasive mechanical ventilation. Obesity 28(7):1195–1199

25. Kompaniyets L, Goodman AB, Belay B, Freedman DS, Sucosky MS, Lange SJ, Gundlapalli AV, Boehmer TK, Blanck HM (2021) Body mass index and risk for COVID-19-related hospitalization, intensive care unit admission, invasive mechanical ventilation, and death-united states, march-december 2020. Morb Mortal Wkly Rep 70(10):355

26. Ranjan P, Kumar A, Chowdhury S, Pandey S, Choudhary A, Bhattacharya A, Singh A, Pandey RM, Wig N, Vikram NK (2020) Is excess weight a risk factor for the development of COVID 19 infection? A preliminary report from India. Diabetes Metab Syndr 14(6):1805–1807

27. Prats-Uribe A, Xie J, Prieto-Alhambra D, Petersen I (2021) Smoking and COVID-19 infection and related mortality: a prospective cohort analysis of UK biobank data. Clin Epidemiol 13:357

28. Patanavanich R, Glantz SA (2020) Smoking is associated with COVID-19 progression: a meta-analysis. Nicotine Tob Res 22(9):1653–1656

29. Desmet, K., Wacziarg, R.: Understanding spatial variation in covid-19 across the united states. nber working paper 27329 (2020)

30. Chhikara BS, Rathi B, Singh J, Poonam F (2020) Corona virus SARS-COV-2 disease COVID-19: infection, prevention and clinical advances of the prospective chemical drug therapeutics. Chem Biol Lett 7(1):63–72

31. Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, Zhang L, Fan G, Xu J, Gu X et al (2020) Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet 395(10223):497–506

32. Chen J, Lu H, Melino G, Boccia S, Piacentini M, Ricciardi W, Wang Y, Shi Y, Zhu T (2020) COVID-19 infection: the China and Italy perspectives. Cell Death Disease 11(6):1–17

33. Chen J, Qi T, Liu L, Ling Y, Qian Z, Li T, Li F, Xu Q, Zhang Y, Xu S et al (2020) Clinical progression of patients with COVID-19 in Shanghai, China. J Infect 80(5):1–6