Heliyon 9 (2023) e19410

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Persistent clinical symptoms and their association with CM syndromes in post-COVID-19 rehabilitation patients in Hong Kong☆

Linda Zhong [a,c,*], Liang Tian [b], Chester Yan Jie Ng [c], Choryin Leung [a], Xian Yang [d], Ching Liong [e], Haiyong Chen [f], Rowena Wong [g], Bacon FL. Ng [g], Z.X. Lin [e], Y.B. Feng [f], Z.X. Bian [a,**], for COVID-19 Research Team

[a] *School of Chinese Medicine, Hong Kong Baptist University, Hong Kong*
[b] *Department of Physics and Institute of Computational and Theoretical Studies, Hong Kong Baptist University, Hong Kong*
[c] *School of Biological Sciences, Nanyang Technological University, Singapore*
[d] *Alliance Manchester Business School, The University of Manchester, Singapore*
[e] *School of Chinese Medicine, The Chinese University of Hong Kong*
[f] *School of Chinese Medicine, LKS Faculty of Medicine, The University of Hong Kong*
[g] *Chinese Medicine Department, Hospital Authority, Hong Kong*

ARTICLE INFO

Keywords:
Chinese medicine
Chinese medicine syndromes
COVID-19
Rehabilitation
Clinical text mining

ABSTRACT

Background: Heterogeneous clinical conditions were observed in individuals who had recovered from COVID-19 and some symptoms were found to persist for an extended period post-COVID. Given the non-specific nature of the symptoms, Chinese medicine (CM) is advantageous in providing holistic medical assessment for individuals experiencing persisting problems. Chinese medicine is a type of treatment that involves prescribing regimens based on CM Syndromes diagnosed by CM practitioners. However, inadequate research on CM elements behind the practice has faced scrutiny.
Methods: This study analysed 1058 CM medical records from 150 post-COVID-19 individuals via a semi-text-mining approach. A logistic model with MCMCglmm was then utilised to analyse the associations between the indicated factors and identified conditions. Calculations were performed using R Studio and related libraries.
Results: With the semi-text-mining approach, three common CM Syndromes (Qi and Yin Deficiency, Lung and Spleen Deficiency, Qi Deficiency of both Spleen and Lung) and nine clinical conditions (fatigue, poor sleep, dry mouth, shortness of breath, cough, headache, tiredness, sweating, coughing phlegm) were identified in the CM clinical records. Analysis via MCMCglmm revealed that the occurrence of persisting clinical conditions was significantly associated with female gender, existing chronic conditions (hypertension, high cholesterol, and diabetes

☆ This research is finally supported by Hong Kong Hospital Authority Research Project: Special Chinese Medicine Out-patient Programme for Discharged Covid-19 Patients – Observational Study (Ref Code: HA 105/48 PT35), Young Qi Huang Scholar Research Grant (SCM-2020-001) and Research Grants Council of Hong Kong (CRF grant no. C2005-22Y).

* Corresponding author. School of Chinese Medicine, Hong Kong Baptist University, Hong Kong.
** Corresponding author.
E-mail addresses: linda.zhong@ntu.edu.sg (L. Zhong), liangtian@hkbu.edu.hk (L. Tian), NG0006ER@e.ntu.edu.sg (C.Y.J. Ng), jasjas@hkbu.edu.hk (C. Leung), xian.yang@manchester.ac.uk (X. Yang), liongching@cuhk.edu.hk (C. Liong), haiyong@hku.hk (H. Chen), rowena.wong@ha.org.hk (R. Wong), ngfl@ha.org.hk (B.FL. Ng), linzx@cuhk.edu.hk (Z.X. Lin), yfeng@hku.hk (Y.B. Feng), bzxiang@hkbu.edu.hk (Z.X. Bian).

https://doi.org/10.1016/j.heliyon.2023.e19410
Received 28 April 2023; Received in revised form 21 August 2023; Accepted 22 August 2023
Available online 25 August 2023

mellitus), and the three persisting CM Syndromes. The current study triangulated the findings from our previous observational study, further showing that patients with certain post-COVID CM Syndromes had significantly increased log-odds of having persisting clinical conditions. Furthermore, this study elucidated that the presence of chronic conditions in the patients would also significantly increase the log-odds of having persistent post-COVID clinical conditions.

Conclusion: This study provided insights on mining text-based CM clinical records to identify persistent post-COVID clinical conditions and the factors associated with their occurrence. Future studies could examine the integration of integrating exercise modules, such as health qigong Liuzijue, into multidisciplinary rehabilitation programmes.

1. Introduction

Chinese medicine (CM), also known as "Traditional Chinese Medicine", is a type of medical practice that focuses on achieving holistic wellbeing of an individual [1]. Its medical theories and treatment regimens are systematically different from that of western medicine (WM), with the former adopting a systemic approach while the latter a reductionist one [2]. With increasing expectations in personalised medicine in recent decades and the bloom of a variety of CM research studies in international journals, CM has been placed under the spotlight outside China [3]. Despite the growth in CM knowledge, the mechanisms behind CM practice have yet to be fully recognised by modern science [4]. CM treatment regimens are generally combinations of Chinese medicines prescribed and dispensed according to CM Syndromes [5], with or without add-on individualised components for different body constitutions identified by the intuition of CM practitioners. Compared with Western medicine, syndrome differentiation is the core of Traditional Chinese medicine by classifying the diseases and selecting the suitable treatment based on each patient's symptoms. More specifically, the symptoms obtained from the four diagnostic methods can be identified into different syndrome types. There are three persistent CM syndromes involved in our findings, which are Qi and Yin Deficiency, Lung and Spleen Deficiency, and Qi Deficiency of both Spleen and Lung. Interestingly, acupuncture, a non-pharmacological type of CM treatment modality [6], is relatively more accepted by the WM community for the management of pain or chronic conditions [7–9], though researchers would continue to seek better-designed clinical trials [10]. Hence, the lack of well-designed modern research studies on the actual CM elements behind the practice contributes to scepticism and rejections in CM [11].

COVID-19 patients have been reported to develop different clinical conditions of variable degrees in different groups [12], and some symptoms were found to persist after hospital discharge [13]. Considering the heterogenic nature of the problem, CM has an advantage in providing a holistic medical perspective for individuals with these health concerns. Our previous study on the effect of individualised CM treatment for COVID-19 rehabilitation [14] identified a change in two CM Syndromes with decreasing clinical symptoms and improved lung functions. In this paper, we analysed CM medical text-based data from the same cohort to elucidate the occurrence of persisting clinical characteristics in these post-COVID patients during their rehabilitation and the potential factors associated with them. We hypothesized that certain CM Syndromes might be associated with these persistent clinical conditions, and these CM Syndromes would likely also occur persistently during rehabilitation despite individualised CM treatment.

This study aims to elucidate the occurrence of persistent clinical symptoms and/or conditions despite individualised CM treatments during COVID-19 rehabilitation via mining CM text-based medical records on CM Syndromes and clinical characteristics, as well as to determine the major factors that might be associated with the occurrence of these persistent clinical conditions, and lastly to form a preliminary decision tree on simple rules for occurrence of persistent clinical conditions.

2. Methods

The multicentre observational study recruited 150 patients who had COVID-19 and were discharged from Hong Kong public hospitals after treatment. They received three to six months of individualised Chinese medicine treatments according to the Chinese Medicine guidelines on COVID-19 rehabilitation. Assessments were made each month and at follow-up on their CM syndromes, BC, lung functions, and other medical conditions.

This study consisted of three parts. The first part of this study utilised a simplified semi-supervised text-mining approach to extract CM Syndromes and clinical conditions from CM clinical text records of the post-COVID-19 patients during rehabilitation. The second part determined the major factors that were associated with the occurrence of persistent clinical conditions via a logistic model with MCMCglmm. The third part constructed a decision tree using the existing data to show simple rules for the occurrence of persistent clinical conditions in post-COVID rehabilitation.

2.1. Subjects and text-based records

We analysed text-based CM medical records from 150 participants who received individualised CM treatments for post-COVID-19 rehabilitation in our previous observational study during the recruitment period between September 7, 2020, to November 30, 2021 [14]. For each participant, a CM medical record was made upon each CM study visit in the observational study (with a maximum of eight visits). In this paper, two main categories of CM medical text-based records were examined: (1) CM Syndromes and (2) clinical characteristics (including clinical symptoms and health conditions). The findings from analysing the text-based records were transformed into factors (occurrence of persistent CM Syndromes and occurrence of persistent clinical conditions) for further analysis with other metadata, such as the demographic characteristics, numbers of days since hospital discharge of the individuals, etc.

2.2. Data processing and analysis

R studio (version 4.1.2, 2021.09.1 Build 372) [15] was used for text-cleaning and analysis with the following libraries installed: “tidyverse” [16], “tidytext” [17], “quanteda” [18], “stringr” [19], “jiebaR” [20], “readtext” [21], “corpus” [22], “tm” [23,24], “tmcn” [25], “ggplot2” [26], “gplots” [27], “dplyr” [28], “wordcloud” [29], “rpart” [30], “rpart.plot” [31], “car” [32], “MASS” [33], and “MCMCglmm” [34].

The two categories of CM medical text-based records were first cleaned to generate two corpora (one for CM Syndromes and one for clinical characteristics). Using R, punctuations and numeric values were stripped from the corpora. In this study, Chinese characters of numbers and numeric attributes such as “days”, “months”, “number of times/counts” were manually removed from the corpora. Additional manual text-cleaning was performed to align differences in Chinese writing styles and expressions with the same meanings. Each finalised corpus was transformed into a term-document matrix (“tdmA” for CM Syndromes and “tdmB” for clinical characteristics as shown in Supplementary Tables 1 and 2). The workflow of text-based data processing is illustrated in Supplementary Fig. 1.

A logistic model with MCMCglmm (“Model”) was constructed to identify factors associated with the occurrence of persistent clinical conditions extracted from the text-based mining in the first part of the study.

The equation of the Model was:

MCMCglmm (Persistent Symptoms ~ Age + Sex + body-mass-index (BMI) Status + Existence of Chronic Conditions + Smoking History + Persistent CM Syndromes + No. of Days from Hospital Discharge, random = ~ Participants + Assigned Data ID, data = Database, prior = Prior, verbose = FALSE).

The equation of Prior for the Model was:

Prior < - list (G = list (G1 = list (V = 1, nu = 0.002), G2 = list (V = 1, nu = 0.002)), R = list (V = 1, nu = 0.002))

To determine whether the Model was affected by the selected Prior, a prior variation was used to check the posterior.mode results. The Prior variation used was:

p.var1 <- var (Database $ Persistent Symptoms, na.rm = TRUE).

Prior1.1 <- list (G = list (G1 = list (V = matrix (p.var1*0.05), nu = 0.002), G2 = list (V = matrix (p.var1*0.05), nu = 0.002)), R = list (V = matrix (p.var1*0.95), nu = 0.002))

Subsequently, logistic regression with MCMCglmm was used to determine the significant factors and obtain the log-odds (post. mean). To verify that the model is not affected by the prior choice, two different priors can be used to perform the logistic regression with MCMCglmm. If no statistical difference is observed between the post.mean calculated with Prior and that calculated with Prior1.1, we can ensure that the model is not affected by the prior choice.

The **paired sample *t*-test** was used to determine the presence of a statistical difference between the two results. Prior to the test, the result of post.mean can be processed according to the significance. Two possibilities are as such: (1) To use the results of post.mean as the input of the paired sample *t*-test directly; or (2) Set the value to zero if the factor is not significant (pMCMC >0.05). For instance, the post.mean of the factor “BMI status – Underweight (score <18.5)” is −0.159, but the p-value is 0.496. Hence, setting such a post.mean value to zero would help avoid possible misleading.

If the second option is chosen, the post.mean calculated with Prior is as shown below. The post.mean of the factor “(Intercept)” is not included, because it is not an actual factor associated with the occurrence of persistent clinical conditions.

$$[0, -0.340, 0, 0, 0, 0, 0.224, 0.209, -0.002]$$

The t statistic is calculated as

$$t = \frac{\overline{X}_D - \mu_0}{s_D / \sqrt{n}}$$

where $\overline{X}_D$ and s_D are the average and standard deviation of the differences between all pairs. The p-value is calculated as

$$P = P(x \leq t)$$

$$p - value = 2 \times \min(P, 1 - P)$$

If the p-value is larger than 0.05, no statistical difference is present between the two results. Paired sample *t*-test can then be conducted using R or python, instead of manual calculation.

In this study, the **mean** of the posterior was used as the result. Similarly, the **median** or the **mode** can also be used as the result, and the paired sample *t*-test can be conducted in the same manner.

Comparison of the posterior.mode (Model $ VCV) of the Model using the above two Priors, little difference was observed (Supplementary Table 3). Hence, the Model was not affected by the Prior choice. A summary of Model was attached as Supplementary Table 4. R generated plots for visual inspection of the Model diagnostics (chain convergence and variance) were shown in Supplementary Figs. 2 and 3.

With the classification and regression tree (CART) algorithm, a preliminary decision tree was created using the existing data (Supplementary Table 5), followed by a precision analysis to determine the usefulness of the tree. To build the training and testing set, the database (with categorisation on occurrence of persistent CM Syndromes and persistent clinical conditions) was scrambled for randomisation and with rows containing missing data removed. The target variable for the decision tree was whether persistent clinical conditions would be present (indicated as "Y" in the leaf of the tree) or not (indicated as "N") in post-COVID individuals possessing certain criteria higher up in the branch.

3. Results

3.1. Longitudinal distribution of medical records

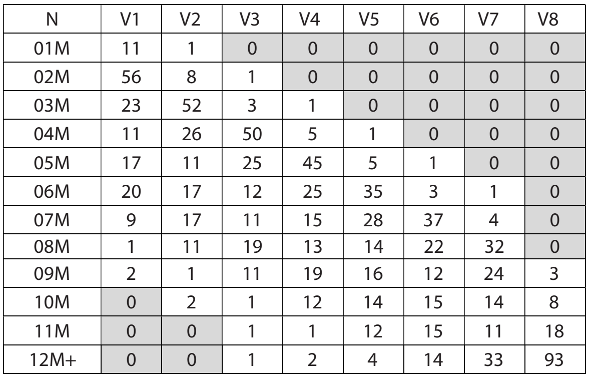
From the 150 participants of the observational study [14], a total of 1058 CM medical records were analysed. The extracted CM medical text records were sorted according to the patients' CM study visit sequence and their post-hospital discharge days (Fig. 1). By taking reference from a post-COVID symptoms integrative classification [35], the time abbreviations refer to the post-discharge time from the hospital. The records in "01 M" are 0–30 days which should be classified as within the "Transition" phase and observations arising within this period could potentially have resulted from the viral infection. Following the classification suggested by Fernández-de-Las-Peñas and colleagues, the "02 M" (31–60 days) to "03 M" (61–90 days) period corresponds to the "Acute" post-COVID phase, "04 M" (91–120 days) to "06 M" (151–180 days) corresponds to the "Long" phase, and "07 M" (181–210 days) onwards corresponds to the "Persistent" phase. However, knowing the patient visit's true length is challenging as each patient has a unique individual medical history and periodicity, which is expressed in medical records as different temporal data. Hence, we have designated time intervals to help structure the follow-up schedule, ensuring that data collection is performed at specific time points consistently, which allows for more systematic and standardized data collection. Most initial CM medical records ("V1") (i.e., the patient's first CM study visit in the observational study) were 31–60 days after discharge from local hospitals ("02 M", i.e., the "Acute" post-COVID phase) (Fig. 1).

3.2. Identification of three persistent CM syndromes and nine persistent clinical conditions

With semi-supervised text-mining approach, three CM Syndromes (Qi and Yin Deficiency, Lung and Spleen Deficiency, Qi Deficiency of both Spleen and Lung) and nine clinical symptoms (fatigue, poor sleep, dry mouth, shortness of breath, cough, headache, tiredness, sweating, coughing phlegm) were found to occur throughout the rehabilitation period with variable expressions in different individuals. The persistence of the elements was determined via a commonality function in the R "wordcloud" library, which identified texts that were found across all the studied time periods. Entry limits were tested at 100, 300, 500, and 1000k word counts respectively, and results generated remained consistent for all tested limits. Wordclouds of common persisting CM Symptoms and clinical conditions were shown in Fig. 2A and B.

Occurrence of persistent clinical conditions associated with female gender, existing chronic conditions, and occurrence of persisting CM Syndromes.

Using the MCMCglmm model, a logistic regression was conducted to determine the significant factors associated with the



N	V1	V2	V3	V4	V5	V6	V7	V8
01M	11	1	0	0	0	0	0	0
02M	56	8	1	0	0	0	0	0
03M	23	52	3	1	0	0	0	0
04M	11	26	50	5	1	0	0	0
05M	17	11	25	45	5	1	0	0
06M	20	17	12	25	35	3	1	0
07M	9	17	11	15	28	37	4	0
08M	1	11	19	13	14	22	32	0
09M	2	1	11	19	16	12	24	3
10M	0	2	1	12	14	15	14	8
11M	0	0	1	1	12	15	11	18
12M+	0	0	1	2	4	14	33	93

Fig. 1. Longitudinal distribution of CM medical records.
A total of 1058 CM medical records from 150 participants of the observational study [14] were analysed in this paper. The medical records spanned a period upon post-COVID hospital discharge (01 M – 12 M+) and further segregated to different CM study visits (V1 – V8). Most initial CM medical records ("V1") (i.e., the patient's first CM study visit in the observational study) were of 31–60 days after discharging from local hospital ("02 M"). Abbreviations for time period post-discharge from hospital: "01 M" – 0 to 30 days; "02 M" – 31 to 60 days; "03 M" – 61 to 90 days; "04 M" – 91 to 120 days; "05 M" – 121 to 150 days; "06 M" – 151 to 180 days; "07 M" – 181 to 210 days; "08 M" – 211 to 240 days; "09 M" – 241 to 270 days; "10 M" – 271 to 300 days; "11 M" – 301 to 330 days; and "12 M+" – over 330 days. For sequence of CM study visits in the observational study [14], abbreviations were as follows: "V1" – initial consultation visit; "V2" – second visit, one month after V1; "V3" – third visit, two months after V1; "V4" – fourth visit, three months after V1; "V5" – fifth visit, four months after V1; "V6" – sixth visit, five months after V1; "V7" – seventh visit, six months after V1; "V8" – follow-up visit, at least three months after patients' last consultation visit.

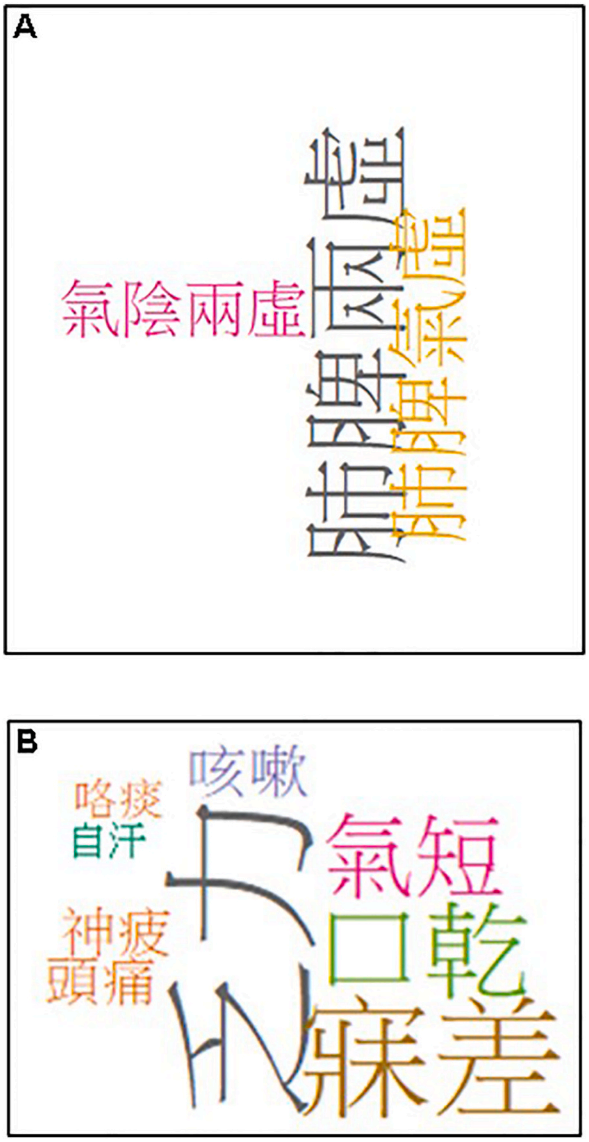


Fig. 2. Wordclouds of frequent and common CM clinical text (in Chinese).
Using R, commonality wordclouds were generated for (A) CM Syndromes and (B) clinical conditions that showed occurrence throughout 01 M–12 M+ during post-COVID rehabilitation. The size of each term shown in the wordcloud indicated its relative frequency among other terms extracted in the analysis. English translations of Chinese terms are provided below for readers' quick reference to the Chinese texts in the wordclouds. For (A) CM Syndromes: Qi and Yin Deficiency [qì yīn liǎng xū, 氣陰兩虛], Lung and Spleen Deficiency [fèi pí liǎng xū, 肺脾兩虛], Qi Deficiency of both Spleen and Lung [fèi pí qì xū, 肺脾氣虛] and (B) clinical symptoms: fatigue [fá lì, 乏力], poor sleep [mèi chà, 寐差], dry mouth [kǒu gān, 口乾], shortness of breath [qì duǎn, 氣短], cough [ké sòu, 咳嗽], headache [tóu tòng, 頭痛], tiredness [shén pí, 神疲], sweating [zì hàn, 自汗], coughing phlegm [kǎ tán, 咯痰].

occurrence of persistent clinical conditions and obtain the log-odds (post.mean). Results showed that the occurrence of persistent clinical conditions was significantly associated with (1) female gender (log-odds of −0.37 for male gender, $p = 0.004$), (2) presence of chronic conditions in the participant (log-odds of 0.24, $p = 0.05$), and interestingly also with (3) the occurrence of any one of the persisting CM Syndromes (log-odds of 0.22, $p = 0.010$) (Table 1). The number of days from hospital discharge (during which the participants received individualised Chinese medicines) was also found to be significantly negatively associated with the occurrence of certain persistent clinical conditions (log-odds of −0.002, $p < 0.001$).

3.3. *Preliminary decision tree analysis*

As shown in Fig. 3, a decision tree was constructed using existing data and the identified persistent CM Syndromes from the first part of the study. A total of 916 records were used for analysis after removing records with missing data. The random allocation of data into a training set (80%) and a testing set (20%) resulted in a similar amount of data when anchored on persistent CM Syndromes (Table 2A).

According to the decision tree (see Fig. 3), the presence of persistent CM Syndromes was shown to be relevant to the occurrence of persistent clinical conditions when the post-COVID individual was over 29 years old and during a 6 to 12-month period after discharge from local hospitals (consisting of 47% of the analysed sample size). Based on this parent node, it was predicted that women would have a 67% chance while men below 44 years old would have a 69% chance of having persistent symptoms if they exhibited any one of the persistent CM Syndromes. On the other hand, splitting from the same parent node, individuals without any of the persistent CM Syndromes were also predicted to experience persistent clinical symptoms if they were over 54 years old and during a 202–251 days period after hospital discharge (64%), or if they were over 68 years old and were over 8-month period post-hospital discharge (86%). However, due to the small sample size of the current analysis, the precision analysis of the decision tree prediction via confusion matrix only reached 64%. As shown in Table 2B, the model misclassified 30 records without any persistent clinical conditions although these records had one or more persistent clinical conditions.

4. Discussion

This paper is the first longitudinal study examining the occurrence of persistent post-COVID clinical conditions in association with CM Syndromes and other factors during rehabilitation. We identified significant factors (female gender, existing chronic conditions, and occurrence of persisting CM Syndromes) associated with the occurrence of one or more of the nine persistent clinical conditions during post-COVID rehabilitation. The occurrence of persistent clinical conditions was also found to be significantly lesser with increasing number of days from hospital discharge. The current study triangulated the findings from our previous observational study and suggested that Chinese medicines alone could improve, but not fully alleviate the residual symptoms.

Post-COVID healthcare has become a global issue following the emergence of a variety of health conditions following COVID-19 recovery [36,37]. These health conditions, also known as "Post-COVID Syndrome" or "COVID-19 long-haul", range from physiological to psychological and are found to vary across individuals [38–41]. While there lies an abundance of studies on post-COVID from the WM point of view [42–44], few practical solutions have been offered and most studies were epidemiological in nature. Despite CM being conventionally used in parallel with WM in China, internationally published randomised controlled trials with individualised CM treatment related to COVID-19 in English were scarce [45–47]. Recently, our group published a prospective observational study, revealing the effects of individualised CM treatment on improving lung functions of post-COVID individuals and diminishing some clinical symptoms though not fully resolving all the clinical conditions [14]. Our current study further provided evidence on significant associations between persisting CM Syndromes and the occurrence of persistent clinical conditions in post-COVID individuals as described by CM practitioners.

The Chinese national guideline "Diagnosis and Treatment Protocol for Novel Coronavirus Pneumonia (Trial Version 9)" [48]

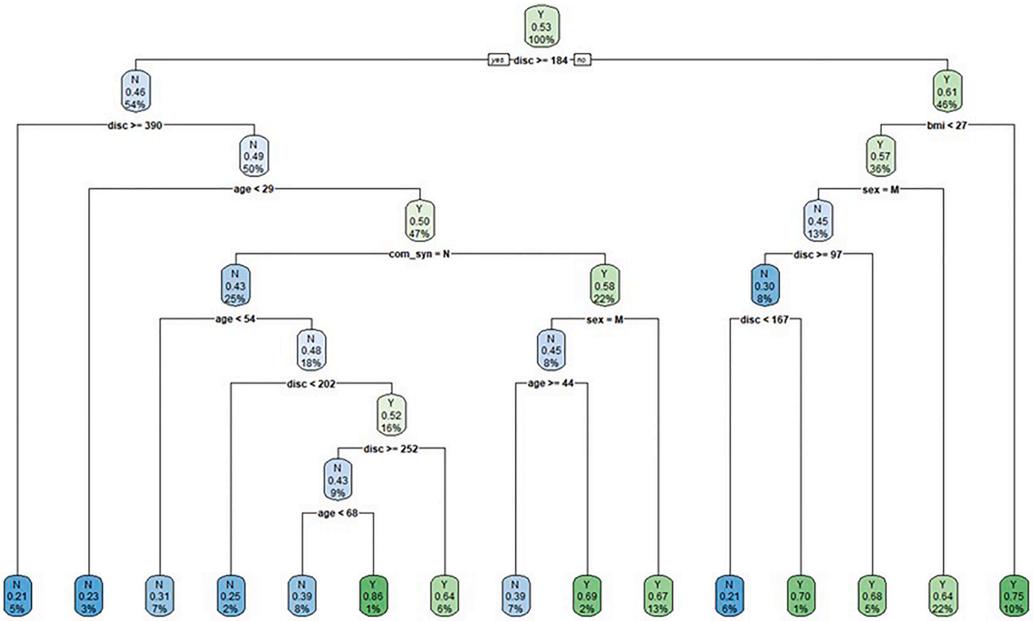


Fig. 3. Decision tree analysis showing simple rules for the occurrence of persistent clinical conditions. Figure showed a decision tree with 64% precision on predicting the occurrence of persistent clinical conditions under various conditions. Abbreviations: sex – Sex; age – Age; bmi – BMI score; com_syn – Persistent CM Syndromes; disc – No. of days from hospital discharge.

Table 1
Summary output of the logistic model with MCMCglmm.

	post.mean	l-95% CI	u-95% CI	eff.samp	pMCMC	Signif.
(Intercept)	**1.0676244**	**0.5999925**	**1.4898346**	**1000**	**<0.001**	***
Age	0.0007715	−0.0082076	0.0083110	1000	0.844	
Sex (Male)	**−0.3674393**	**−0.5886256**	**−0.1473285**	**1135**	**0.004**	**
BMI status – Obese (score 25–29.9)	0.0380184	−0.2444710	0.2948440	1000	0.790	
BMI status – Overweight (score 23–24.9)	0.0369740	−0.2073175	0.2960266	1000	0.782	
BMI status – Underweight (score <18.5)	−0.1566723	−0.5927793	0.2974664	1000	0.506	
BMI status – Very obese (score ≥30)	−0.0329520	−0.3721664	0.2728272	1000	0.856	
Chronic conditions (With)	**0.2352174**	**−0.0185709**	**0.4264830**	**1112**	**0.05**	.
Smoking history (With)	0.0828884	−0.2296322	0.3514871	1000	0.574	
Persistent CM Syndromes (With)	**0.2221648**	**0.0483805**	**0.3916519**	**1199**	**0.010**	*
No. of days from hospital discharge	**−0.0021638**	**−0.0027886**	**−0.0016281**	**1000**	**<0.001**	***

Table 2A
Quality of decision tree model (A) Randomisation of data into training set and testing set using the persistent CM Syndromes factor as the anchor.

	N	Y
Training set	0.54	0.46
Testing set	0.53	0.47

N – Do not have any of the persistent CM Syndromes.
Y – Have at least one of the persistent CM Syndromes.

Table 2B
Confusion matrix on precision analysis of decision tree prediction.

		Predicted	
		N	Y
Actual	**N**	37	36
	Y	30	81

N – Do not have any of the persistent clinical conditions.
Y – Have at least one of the persistent clinical conditions.

documents the clinical experiences accumulated by both CM practitioners and WM doctors at the pandemic frontline in China. Clinical conditions identified in the COVID-19 rehabilitation phase were general shortness of breath, fatigue or tiredness, poor appetite, nausea, vomiting, stomach fullness, difficulty in defecation, and watery stool for individuals with the CM Syndrome "Qi Deficiency of both Spleen and Lung"; while general fatigue, shortness of breath, dry mouth, feeling thirsty, palpitation, profuse sweating, poor appetite, mild or no fever, and dry cough with little sputum were observed for those with CM Syndrome "Qi and Yin Deficiency" [48]. Among the above clinical conditions, semi-text-mining of CM medical records identified shortness of breath, fatigue, tiredness, dry mouth, and cough as common persistent conditions suffered by post-COVID individuals in Hong Kong despite the prescription of individualised CM treatments, and the occurrence of these persistent clinical symptoms was associated with the above reported CM Syndromes. Additional persistent clinical conditions observed in post-COVID individuals in the current study included poor sleep, headache, sweating, and coughing of phlegm (Fig. 2).

The female gender has been reported to be associated with a higher risk of persistent symptoms such as chronic fatigue and respiratory problems [49,50]. Our findings concurred with other studies that the male gender was at lower odds with the occurrence of persistent clinical conditions. We further noticed from our preliminary decision tree that under certain criteria where gender constituted a critical split, females had a 64–67% chance of having one or more persisting clinical conditions compared to 45% in males (Fig. 3). This observation remained true regardless of the presence of persisting CM Syndromes. However, this observation should be interpreted with caution as the sample size was small, and the precision of the decision tree was not optimal. Contrary to previous studies by others, our paper did not find any significant associations between the occurrence of persistent clinical conditions, age [50,51] and BMI of individuals [50,52,53]. The number of days from hospital discharge was found to be negatively associated with the occurrence of persistent clinical conditions (Table 1). This observation might be specific for the current study set in which individuals had been receiving Chinese medicines prescriptions during the observational study that would improve their clinical conditions. Another hypothesis might be, given a longer period post-COVID, the "persistent" clinical conditions would resolve with or without treatments. A larger sample size, or even a clinical trial with a non-CM control group, would be necessary to validate this observation.

For our study, we have made use of semi-supervised text mining for our data analysis due to the following advantages. Firstly, semi-supervised methods have limited reliance on labelled data. Obtaining a significant number of precisely labelled data in the medical profession can be difficult due to variables such as privacy issues, time-consuming manual annotation methods, and the requirement for subject expertise. Semi-supervised methods use fewer labelled records than fully supervised methods, thus making them more viable and cost-effective in circumstances when labelled data is limited. Additionally, the use of semi-supervised methods enhances the generalization of unlabelled data, which is greatly beneficial when mining medical records, which tend to contain a large amount of relatively easily obtainable unlabelled data. By incorporating unlabelled examples during training, which provides valuable information about data distribution and structure, the model can learn more representative features and handle the variations and complexities present in medical records. This results in improved generalization and better performance when faced with new and unseen data. Lastly, semi-supervised methods are flexible and adaptable to different medical domains and tasks. They allow for the incorporation of domain-specific knowledge and can effectively handle variations in terminology, document structure, and data sources. This adaptability is crucial in the medical field as reporting practices may vary among different specialties, healthcare centres, or countries. As for simplicity, the complexity of semi-supervised methods depends on the specific algorithm and implementation used. While the overall framework of semi-supervised learning may introduce additional complexity compared to supervised learning, various techniques can simplify the process.

Our study also provided evidence, based on a semi-supervised text-mining approach, to triangulate the identification of the two main CM Syndromes (Qi and Yin Deficiency and Qi Deficiency of both Spleen and Lung) during COVID-19 rehabilitation phase by the Chinese national guideline [48]. The current text-mining findings concurred with our previously published results using quantitative CM Syndrome assessments, confirming the deficiency in Qi and Yin, as well as Spleen and Lung deficiency, as two major CM Syndromes during post-COVID rehabilitation [14]. CM practitioners generally record clinical observations in a text-based manner. Such clinical text may include but is not limited to: CM Syndromes, clinical features, and symptoms, living habits, and other observations as deemed relevant to the diagnosis at the time of consultation. These clinical texts recorded by CM practitioners would provide useful data to understand and manage patient heterogeneity. Currently, there is no “gold” standard of corpora for CM that could be adopted for use in analysing CM medical text data [54]. The unstructured nature and variable forms of Chinese text expressions makes data processing more difficult and hence requires additional manpower [54]. In the current study, the writing styles and habitual vocabularies were relatively less complex compared to other full text-mining analysis of complete CM medical records since this study only focused on CM Syndromes and clinical characteristics as recorded in the case report files of the previous observational study [14].

With rapid advancements in computer science, network pharmacology has been increasingly used to study the complex interactions between Chinese medicines (whether in combinations as a Formula or as a single ingredient) and biological functions or molecular mechanisms [55]. However, such studies rarely include CM clinical text-based information and are often designed in a cross-sectional manner. CM treatment regimens are prescribed according to CM theories which are determined by different types of CM Syndromes that are highly dynamic and changing between treatments. Studies investigating CM clinical text elements are usually published in Chinese journals with investigators presenting their findings in a descriptive manner. Without a systematic approach to interpret the CM text-based data in CM clinical research, valuable CM clinical experience which forms the foundation for CM practice would remain elusive and extremely difficult for professionals of other medical specialties to review the CM findings. Here, we also call for future research to establish a Chinese medicine corpus as the gold standard for natural language processing tasks in the field of Chinese medicine. By constructing a comprehensive Chinese medicine corpus, knowledge from a wide range of traditional Chinese medicine texts, such as ancient medical classics, clinical records, herbology, acupuncture, and traditional prescriptions, can be organized, digitized, and made accessible for analysis. This facilitates knowledge discovery, data mining, and integration of information from various sources. This standardization enhances collaboration, data sharing, and reproducibility of research in the field of Chinese medicine. The current study utilised the rich CM clinical experience documented in the text-based medical records to determine the factors associated with the occurrence of persistent clinical conditions.

Despite individualised Chinese medicines treatment received by the patients in this study, the residual conditions persisted for a long period of time. The fact that no single rehabilitation regime has been reported to fully resolve these problems leads us to postulate that an integrative rehabilitation approach might be necessary. Apart from conventional WM and individualised CM, exercise has been reported to improve aspects of the immune system [56,57], fatigue [58], and physical fitness encompassing muscle strength and endurance [59]. In the CM medical system, exercise routines such as Liuzijue and Baduanjin are commonly used for maintaining regular health instead of “treatment”. They encompass both slow and mild physical exercise with breathing regulation, as well as the mental aspect of “peacefulness”. These simple routines are home exercise programs and are generally easy to learn, thus making them highly accessible to the public. A recent study by Tang and colleagues reported that Liuzijue could reduce breathlessness and improve the mental states in discharged COVID-19 patients [60]. Exercise has also been suggested to improve the functional reserve of the body and would be beneficial for post-COVID recovery [61]. It would be interesting to elucidate whether a combination of different approaches would eliminate the post-COVID clinical conditions.

In summary, we identified an association between the occurrence of persisting CM Syndromes, the female gender, and the existence of chronic conditions with the occurrence of persistent clinical conditions in post-COVID individuals. Indeed, this study has certain limitations. Firstly, because of the current small sample size, the decision tree prediction model was not helpful in devising further strategies on combating the persistent clinical problems. Analysis of a larger set of data would be required to increase the precision of the decision tree prediction. Secondly, as mentioned previously, CM encompasses a vast body of knowledge with diverse terminology and concepts and the lack of standardized terminology poses challenges for text mining techniques that heavily rely on consistent and well-defined language. Moreover, CM concepts and theories are deeply rooted in Chinese culture and philosophy, requiring a deep understanding of the context to correctly interpret and extract meaningful information. Furthermore, the performance of the semi-supervised approach used in this study largely depends on the quality of the data. If the data contains noise, inconsistencies, or irrelevant information, it may adversely affect the model’s performance. Additionally, evaluating the quality and reliability of information extracted from TCM texts using text mining methods may be difficult, leading to biased. Despite the limitation of the current study, we hope that light would be shed on future integrative research methods to elucidate CM in its entirety, encompassing both qualitative and quantitative aspects of the CM practice in the research design.

5. Conclusions

In conclusion, we provided evidence of the factors that were associated with the occurrence of persistent clinical conditions during post-COVID rehabilitation. This study provided insights into using text-based CM clinical records to identify persistent post-COVID clinical conditions and the factors associated with their occurrence. By integrating various modern research methodologies, CM practitioners and researchers are empowered to elucidate complex and dynamic clinical conditions, using both objective assessments and evidence-based evaluated subjective CM clinical experience to contribute useful information to the scientific community. Furthermore, future studies could explore integrating exercise modules, such as Liuzijue, into a multidisciplinary rehabilitation program.

[item-group: IG000074].

Author contribution statement

Linda Zhong: Conceived and designed the experiments, Performed the experiments, Analysed and interpreted the data, Contributed reagents, materials, analysis tools or data, Wrote the paper.

Liang Tian, Chester Yan Jie Ng, Xian Yang: Analysed and interpreted the data, Wrote the paper.

Choryin Leung, Ching Liong, Haiyong Chen: Performed the experiments, Wrote the paper.

Rowena Wong, Bacon FL Ng: Contributed reagents, materials, analysis tools or data; Wrote the paper.

ZX Lin, YB Feng, ZX Bian: Conceived and designed the experiments, Wrote the paper.

Data availability statement

Data included in article/supp. material/referenced in article.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2023.e19410.

N	V1	V2	V3	V4	V5	V6	V7	V8
01M	11	1	0	0	0	0	0	0
02M	56	8	1	0	0	0	0	0
03M	23	52	3	1	0	0	0	0
04M	11	26	50	5	1	0	0	0
05M	17	11	25	45	5	1	0	0
06M	20	17	12	25	35	3	1	0
07M	9	17	11	15	28	37	4	0
08M	1	11	19	13	14	22	32	0
09M	2	1	11	19	16	12	24	3
10M	0	2	1	12	14	15	14	8
11M	0	0	1	1	12	15	11	18
12M+	0	0	1	2	4	14	33	93

References

[1] K. Shankar, L.P. Liao, Traditional systems of medicine, Phys Med Rehabil Clin N Am 15 (4) (2004) 725–747.

[2] J. van der Greef, et al., Systems biology-based diagnostic principles as pillars of the bridge between Chinese and Western medicine, Planta Med. 76 (17) (2010) 2036–2047.

[3] H. Uzuner, et al., Traditional Chinese medicine research in the post-genomic era: good practice, priorities, challenges and opportunities, J. Ethnopharmacol. 140 (3) (2012) 458–468.

[4] W.Y. Jiang, Therapeutic wisdom in traditional Chinese medicine: a perspective from modern science, Discov. Med. 5 (29) (2005) 455–461.

[5] F. Cheng, et al., Biologic basis of TCM syndromes and the standardization of syndrome classification, J Traditional Chinese Med Sci 1 (2) (2014) 92–97.

[6] W. Zhou, P. Benharash, Effects and mechanisms of acupuncture based on the principle of meridians, J Acupunct Meridian Stud 7 (4) (2014) 190–193.

[7] A.A. Berger, et al., Efficacy of acupuncture in the treatment of chronic abdominal pain, Anesthesiol. Pain Med. 11 (2) (2021), e113027.
[8] H. Foley, A. Steel, J. Adams, Perceptions of person-centred care amongst individuals with chronic conditions who consult complementary medicine practitioners, Complement Ther Med 52 (2020), 102518.
[9] M. Patel, et al., The role of acupuncture in the treatment of chronic pain, Best Pract. Res. Clin. Anaesthesiol. 34 (3) (2020) 603–616.
[10] I. Urits, et al., A comprehensive review of alternative therapies for the management of chronic pain patients: acupuncture, tai chi, osteopathic manipulative medicine, and chiropractic care, Adv. Ther. 38 (1) (2021) 76–89.
[11] D. Cyranoski, Why Chinese medicine is heading for clinics around the world, Nature 561 (7724) (2018) 448–450.
[12] W.J. Guan, et al., Clinical characteristics of coronavirus disease 2019 in China, N. Engl. J. Med. 382 (18) (2020) 1708–1720.
[13] T. Yang, et al., Sequelae of COVID-19 among previously hospitalized patients up to 1 year after discharge: a systematic review and meta-analysis, Infection 50 (5) (2022) 1067–1109.
[14] L.L. Zhong, et al., Effects of Chinese medicine for COVID-19 rehabilitation: a multicenter observational study, Chin. Med. 17 (1) (2022) 99.
[15] R. Team, RStudio: Integrated Development for R, RStudio, PBC, Boston, MA, 2020.
[16] H. Wickham, et al., Welcome to the tidyverse, J. Open Source Softw. 4 (43) (2019).
[17] J. Silge, D. Robinson, Tidytext: text mining and analysis using tidy data principles in R, J. Open Source Softw. 1 (3) (2016).
[18] K. Benoit, et al., quanteda: an R package for the quantitative analysis of textual data, J. Open Source Softw. 3 (30) (2018).
[19] H stringr Wickham, Simple, Consistent Wrappers for Common String Operations, 2022.
[20] W. Qin, Y. Wu jiebaR, Chinese Text Segmentation, 2019.
[21] K. Benoit, A. Obeng readtext, Import and Handling for Plain and Formatted Text Files, 2021.
[22] P.O. Perry, Corpus, Text Corpus Analysis, 2021.
[23] I. Feinerer, K. Hornik tm, Text Mining Package, 2020.
[24] I. Feinerer, K. Hornik, D. Meyer, Text mining infrastructure in R, J. Stat. Software 25 (5) (2008).
[25] J. tmcn Li, A Text Mining Toolkit for Chinese, 2019.
[26] H. Wickham, ggplot2: Elegant Graphics for Data Analysis, Springer-Verlag, New York, 2016.
[27] G.R. Warnes, et al., Gplots: Various R Programming Tools for Plotting Data, 2022.
[28] H. Wickham, et al., Dplyr: A Grammar of Data Manipulation, 2021.
[29] I. wordcloud Fellows, Word Clouds, 2018.
[30] T. Therneau, B. Atkinson rpart, Recursive Partitioning and Regression Trees, 2022.
[31] S. Milborrow, rpart.plot: Plot 'rpart' Models: an Enhanced Version of 'plot, rpart, 2022.
[32] J. Fox, S. Weisberg, An {R} Companion to Applied Regression, 2019.
[33] W.N. Venables, B.D. Ripley, Modern Applied Statistics with S, Springer, 2002.
[34] J.D. Hadfield, MCMC methods for multi-response generalized linear mixed models: TheMCMCglmmRPackage, J. Stat. Software 33 (2) (2010).
[35] C. Fernandez-de-Las-Penas, et al., Defining post-COVID symptoms (Post-Acute COVID, long COVID, persistent post-COVID): an integrative classification, Int J Environ Res Public Health 18 (5) (2021).
[36] T. Greenhalgh, et al., Management of post-acute covid-19 in primary care, BMJ 370 (2020) m3026.
[37] D.L. Sykes, et al., Post-COVID-19 symptom burden: what is long-COVID and how should we manage it? Lung 199 (2) (2021) 113–119.
[38] H. Crook, et al., Long covid-mechanisms, risk factors, and management, BMJ 374 (2021) n1648.
[39] E.P. Hoffer, Long COVID: does it exist? What is it? We can we Do for sufferers? Am. J. Med. 134 (11) (2021) 1310–1311.
[40] Y. Huang, et al., COVID symptoms, symptom clusters, and predictors for becoming a long-hauler looking for clarity in the haze of the pandemic, Clin. Nurs. Res. 31 (8) (2022) 1390–1398.
[41] G. Vanichkachorn, et al., Post-COVID-19 syndrome (long haul syndrome): description of a multidisciplinary clinic at mayo clinic and characteristics of the initial patient cohort, Mayo Clin. Proc. 96 (7) (2021) 1782–1791.
[42] C. Huang, et al., 6-month consequences of COVID-19 in patients discharged from hospital: a cohort study, Lancet 397 (10270) (2021) 220–232.
[43] A. Nalbandian, et al., Post-acute COVID-19 syndrome, Nat Med 27 (4) (2021) 601–615.
[44] Z. Yan, M. Yang, C.L. Lai, Long COVID-19 syndrome: a comprehensive review of its effect on various organ systems and recommendation on rehabilitation plans, Biomedicines 9 (8) (2021).
[45] X. Wenguang, et al., Randomized controlled study of a diagnosis and treatment plan for moderate coronavirus disease 2019 that integrates Traditional Chinese and Western Medicine, J. Tradit. Chin. Med. 42 (2) (2022) 234–241.
[46] Y.A. Ye, G.C.C. Group, Guideline-based Chinese herbal medicine treatment plus standard care for severe coronavirus disease 2019 (G-champs): evidence from China, Front. Med. 7 (2020) 256.
[47] J. Zhao, et al., Yidu-toxicity blocking lung decoction ameliorates inflammation in severe pneumonia of SARS-COV-2 patients with Yidu-toxicity blocking lung syndrome by eliminating IL-6 and TNF-a, Biomed. Pharmacother. 129 (2020), 110436.
[48] NHC, NATCM, Diagnosis and Treatment Protocol for Novel Coronavirus Pneumonia, ninth ed., 2022.
[49] D. Munblit, et al., Incidence and risk factors for persistent symptoms in adults previously hospitalized for COVID-19, Clin. Exp. Allergy 51 (9) (2021) 1107–1120.
[50] C.H. Sudre, et al., Attributes and predictors of long COVID, Nat Med 27 (4) (2021) 626–631.
[51] S. Damanti, et al., Prevalence of long COVID-19 symptoms after hospital discharge in frail and robust patients, Front. Med. 9 (2022), 834887.
[52] S.H. Loosen, et al., Obesity and lipid metabolism disorders determine the risk for development of long COVID syndrome: a cross-sectional study from 50,402 COVID-19 patients, Infection 50 (5) (2022) 1165–1170.
[53] R.S. Peter, A. Nieters, S.O. Brockmann, S. Göpel, G. Kindle, U. Merle, J.M. Steinacker, W.V. Kern, D. Rothenbacher, EPILOC Phase 1 Study Group, Association of BMI with general health, working capacity recovered, and post-acute sequelae of COVID-19, Obesity 31 (1) (2023) 43–48.
[54] T. Zhang, et al., Information extraction from the text data on traditional Chinese medicine: a review on tasks, challenges, and methods from 2010 to 2021, Evid Based Complement Alternat Med 2022 (2022), 1679589.
[55] X. Wang, et al., TCM network pharmacology: a new trend towards combining computational, experimental and clinical approaches, Chin. J. Nat. Med. 19 (1) (2021) 1–11.
[56] R.M. Barker-Davies, et al., The Stanford Hall consensus statement for post-COVID-19 rehabilitation, Br. J. Sports Med. 54 (16) (2020) 949–959.
[57] D.C. Nieman, Exercise is medicine for immune function: implication for COVID-19, Curr. Sports Med. Rep. 20 (8) (2021) 395–401.
[58] E. Daynes, et al., Early experiences of rehabilitation for individuals post-COVID to improve fatigue, breathlessness exercise capacity and cognition - a cohort study, Chron. Respir. Dis. 18 (2021), 14799731211015691.
[59] H. Chaabene, et al., Home-based exercise programmes improve physical fitness of healthy older adults: a PRISMA-compliant systematic review and meta-analysis with relevance for COVID-19, Ageing Res. Rev. 67 (2021), 101265.
[60] Y. Tang, et al., Liuzijue is a promising exercise option for rehabilitating discharged COVID-19 patients, Medicine (Baltim.) 100 (6) (2021), e24564.
[61] D. Jimenez-Pavon, A. Carbonell-Baeza, C.J. Lavie, Physical exercise as therapy to fight against the mental and physical consequences of COVID-19 quarantine: special focus in older people, Prog. Cardiovasc. Dis. 63 (3) (2020) 386–388.