Journal of Infection and Public Health 12 (2019) 700–704

Contents lists available at ScienceDirect

Journal of Infection and Public Health

journal homepage: http://www.elsevier.com/locate/jiph

Main factors influencing recovery in MERS Co-V patients using machine learning

Maya John [1,*], Hadil Shaiba [2]

[1] Department of Computer Science and Engineering, Sree Buddha College of Engineering, Pathanamthitta, Kerala, India
[2] Department of Computer Science, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia

ARTICLE INFO

Article history:
Received 13 November 2018
Received in revised form 8 March 2019
Accepted 24 March 2019

Keywords:
MERS
Infectious disease
Survival rate
Machine learning
Saudi Arabia

ABSTRACT

Background: Middle East Respiratory Syndrome (MERS) is a major infectious disease which has affected the Middle Eastern countries, especially the Kingdom of Saudi Arabia (KSA) since 2012. The high mortality rate associated with this disease has been a major cause of concern. This paper aims at identifying the major factors influencing MERS recovery in KSA.
Methods: The data used for analysis was collected from the Ministry of Health website, KSA. The important factors impelling the recovery are found using machine learning. Machine learning models such as support vector machine, conditional inference tree, naïve Bayes and J48 are modelled to identify the important factors. Univariate and multivariate logistic regression analysis is also carried out to identify the significant factors statistically.
Result: The main factors influencing MERS recovery rate are identified as age, pre-existing diseases, severity of disease and whether the patient is a healthcare worker or not. In spite of MERS being a zoonotic disease, contact with camels is not a major factor influencing recovery.
Conclusion: The methods used were able to determine the prime factors influencing MERS recovery. It can be comprehended that awareness about symptoms and seeking medical intervention at the onset of development of symptoms will make a long way in reducing the mortality rate.

Introduction

The devastating marks left by infectious diseases on human race are much similar to that caused by famine and war [1]. Majority of the infectious diseases are caused by viruses, bacteria or fungi. Coronaviruses are a family of viruses responsible for causing diseases ranging from mild infection to death causing diseases such as Severe Acute Respiratory Syndrome (SARS). Since the advent of the 21st century, there has been a surge in the discovery of new coronaviruses such as SARS Co-V, HCoV-NL63, HCoV-HKU1 etc. [2]. The detailed medical analysis on the death of a 60-year-old man in Saudi Arabia led to the discovery of a novel coronavirus [2], which was later coined as Middle East Respiratory Syndrome CoronaVirus (MERS Co-V). The incubation period associated with the viral infection ranges from 2 to 14 days. The mortality rate associated with MERS Co-V is over 30% [3]. MERS Co-V is a zoonotic virus wherein human beings may get infected due to direct or indirect contact with infected animals [4]. Camels have been identified as a carrier [5]. It is believed that the MERS-CoV is transmitted to humans from camels through respiratory droplet, saliva, or consuming undercooked camel meat or milk. Using viral genome sequencing, the zoonotic nature of the virus was confirmed [6]. MERS Co-V RNA was detected in nasal swab of infected camels, and their antibodies were found in 70% of the positively tested camels [7]. Contact with infected people in healthcare facilities, houses and communities have also contributed to the spread of the disease [8–11]. Healthcare facilities are the major source of nosocomial outbreak of the disease [12]. Studies conducted have reported that there is no pandemic risk associated with the disease acquired through contact with infected people [13]. Proper awareness of modes of transmission of the disease can help in preventing the spread of the infection. Several studies have been conducted to assess the knowledge of MERS among various sections of the society [14–16]. A recent study conducted throws light on the lack of knowledge of the general public regarding the probable epidemic nature of the disease [15].

* Corresponding author.
E-mail addresses: maya.j.mail@gmail.com (M. John), hadil.shaiba@gmail.com (H. Shaiba).

https://doi.org/10.1016/j.jiph.2019.03.020

Table 1
Description of data used.

Attribute	Description of attribute	Values	% of cases	Death	Recovery
Gender	Gender of the patient	Male	69.6%	38.3%	61.7%
		Female	30.4%	21.8%	78.2%
Age	Age of the patient	1–25 years	5.9%	21.7%	78.3%
		26–50 years	39.2%	18.9%	81.1%
		51–75 years	44.5%	41.3%	58.7%
		>75 years	10.4%	59.8%	40.2%
HCW	Healthcare worker or not	Yes	16.1%	0.8%	99.2%
		No	83.9%	39.5%	60.5%
Symptoms	Symptoms present or not	Yes	89.8%	37.1%	62.9%
		No	10.2%	0%	100%
IStatus	Status at time of identification of disease	Stable	65.3%	17.4%	82.6%
		Critical	34.7%	63.2%	36.8%
PreDisease	Presence of pre-existing disease or not	Yes	71.8%	43%	57%
		No	28.2%	8.6%	91.3%
AnimalExp	Patient in contact with animal or not	Yes	21.2%	37.9%	62.1%
		No	78.8%	32%	68%
HHC	Hospital, household or community acquired	Yes	38.3%	21.3%	78.7%
		No	53.2%	39.3%	60.7%
		Under investigation	8.5%	49.2%	50.8%
DR	Patient died or recovered	Died	33.3%	–	–
		Recovered	66.7%	–	–

According to the information provided by the World Health Organization (WHO), about 27 countries have reported cases on MERS Co-V infection since 2012. Nearly 80% of the cases have been reported from the Middle Eastern country, Kingdom of Saudi Arabia. Acute respiratory problem along with fever, cough or shortness of breath are the most common symptoms of MERS Co-V infection. Apart from the aforementioned health problems, many people infected with the virus suffered from diarrhoea, nausea or vomiting. In a small fraction of cases, people with infection showed very mild or no symptoms. The mortality rate is comparatively high in people suffering from comorbidities [17]. People suffering from kidney problems, cancer, diabetics, respiratory problems etc. are more likely to be severely infected by the virus [12].

Rivers et al. used Poisson regression to identify the risk factors associated with death and severity of the disease. They concluded that advancing age and pre-existing diseases were the main risk factors leading to death and severe infection [18]. Ahmed employed statistical methods to determine the predictors responsible for recovery from the disease after 3 and 30 days of onset of symptoms. The study identified old age, non-healthcare workers, severe infection and hospital-acquired infection as the leading factors responsible for mortality [19].

Al-Turaiki et al. used predictive models such as naïve Bayes and J48 to analyse MERS data pertaining to Saudi Arabia [20]. The recovery model generated indicates that the healthcare workers are most likely to recover from the infection. In the case of non-healthcare workers, pre-existing diseases hamper the recovery. The authors also generated a stability model from which it can be inferred that in the case of people with pre-existing diseases, symptoms and age turned out to be the significant factors in determining the severity of viral infection. Abdullah et. al. used classification methods such as support vector machine, random forest and naïve Bayes to detect MERS Co-V infection. The authors observed that people above 50 years had more chances of infection compared to the rest [21].

The studies on the trends associated with infectious diseases are broadly classified into computational, mathematical and surveillance based methods [22]. Our paper aims at identifying the important factors which influence the recovery of MERS Co-V infected people in the Kingdom of Saudi Arabia. Machine learning techniques and statistical analysis are employed for the purpose.

Methods

Data description

The data for analysis was taken from the Control and Command Centre, Ministry of Health website of the Kingdom of Saudi Arabia. MERS Co-V cases from January 2015 to April 2018 have been used in the study. A total number of 983 MERS cases were reported during this period. Although there has been a tremendous reduction in the number of new cases compared to the year 2014 until now the disease has not been brought under control.

The information gathered consisted of daily records based on new cases, recovery cases and death cases. The recovery and death cases were mapped with the new cases reported to gather more information regarding the patients infected with MERS. The unmatched patient records were eliminated and hence 836 patient records were used for analysis. Out of the 836 cases, 52 patients were initially reported as dead. Hence, those cases were also removed from the dataset, and 784 cases were used in the study. The description of the database used for research is shown in Table 1. The database analysed consists of 261 death cases.

Methods used

Machine learning methods were used to find the important factors associated with viral infections. The experiments were implemented using R programming language. The machine learning methods such as support vector machine, naïve Bayes, conditional inference tree, J48 and logistic regression were used to find the important predictors. The variable importance associated with the factors serves as an indicator of the importance of the factors associated with the recovery from MERS Co-V infection. Univariate and multivariate analysis were carried out using logistic regression to identify the statistically significant factors.

Support vector machine

Support vector machine is also known as hyperplane classifier as it strives to find a hyperplane which is capable of dividing the different classes of data with a relatively large margin. It has the capability to learn, independent of the number of features of the dataset. SVM with a radial kernel is used to perform the experiments.

Table 2
Univariate and multivariate logistic regression analysis of recovery.

Characteristic	Univariate analysis		Multivariate analysis	
	Estimate	p-Value	Estimate	p-Value
Gender (male)	−0.7968	9.46e-06 ***	−0.40129	0.07732.
Age	−0.8332	1.07e-13 ***	−0.43841	0.00135 **
Healthcare worker (yes)	4.40252	1.23e-05 ***	3.16166	0.00233 **
Symptoms (yes)	−17.04	0.969	−16.03475	0.97933
Initial status (stable)	2.1011	<2e-16 ***	1.79366	<2e-16 ***
Pre-existing disease (yes)	−2.0813	2.97e-16 ***	−0.96765	0.00137 **
Animal exposure (yes)	−0.26039	0.152	0.08942	0.69808
Hospital household community acquired (under investigation)	−0.4037	0.126	−0.65328	0.04246 *
Hospital household community acquired (yes)	0.8714	4.69e-07 ***	−0.19602	0.40661

Naives Bayes

Naives Bayes is based on the assumption that the factors of the dataset considered are in no way related to one another. It is based on Bayes' theorem which computes the conditional probability of the classification result based on independent factors [23]. Conditional probability is computed as shown in Eq. (1).

$$P(C|X) = \frac{P(X|C)^{*}P(C)}{P(X)} \quad (1)$$

where C denotes the target variable and X is the predictor.

Conditional inference tree

Conditional inference tree is a widely used tree-based classification technique. It differs from RPart in the strategy adopted to select variables while splitting the tree. RPart selects variable by maximizing information criterion while conditional inference tree selects variables which are statistically significant.

J48

J48 generates C4.5 trees which are either pruned or unpruned [24]. The decision tree method C4.5 is an extension of Iterative Dichotomiser 3 (ID3). The nodes are split based on the difference in entropy. C4.5 is capable of handling both continuous and discrete attributes.

Logistic regression

Logistic regression (also known as logit regression) is a classification method used to predict the value of a categorical variable. It models the probability of the response belonging to a particular category. Logistic regression involves computing the log odds of an event which is considered equivalent to multiple linear regression function. In this work, binomial logistic regression was used because the outcome of the prediction belongs to any one of the two classes (death or recovery).

Results

The important predictors associated with MERS recovery was obtained by computing the variable importance associated with different machine learning methods by considering the entire data as the train data. The graphical representation of the variable importance of factors corresponding to various methods are shown in Fig. 1.

It is evident from various plots in Fig. 1 that the four most important factors associated with survival of MERS Co-V infected people are initial status of the disease (istatus), age, pre-existing diseases (predisease) and whether the patient is a healthcare worker or not (hcw). Statistical analysis such as univariate and multivariate logistic regression were also carried out to find the important factors associated with recovery from MERS. The estimate and p-values of the analysis are tabulated in Table 2.

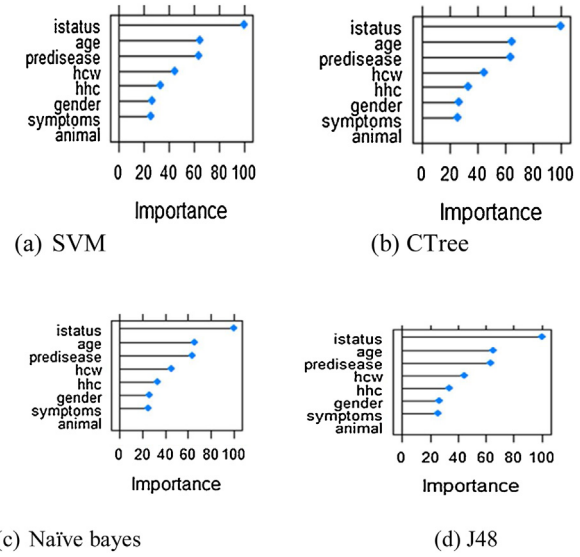

Fig. 1. Graphical representation of variable importance corresponding to different methods.

Table 3
Multivariate logistic regression analysis of recovery using top 4 significant variables.

Characteristic	Estimate	p-Value
Age	−0.4327	0.001278 **
Healthcare worker (yes)	3.2987	0.001260 **
Initial status (stable)	1.8837	< 2e-16 ***
Pre-existing disease (yes)	−0.9861	0.001067 **

The presence of asterisk next to the p-value in Table 2 indicates that the attribute is significant. The p-values of significant variables are visually encoded. The significance codes less than 0.001, between 0.001 and 0.01, between 0.01 and 0.05, and between 0.05 and 0.1 are represented using the symbol ***, **, * and . (dot) respectively. In the aforementioned table, estimate refers to the log of odds. A negative value of estimate indicates that the variable has a negative impact on recovery and vice versa. It can be inferred from univariate and multivariate analysis that all variables except symptoms and animal exposure are significant in determining the recovery of a patient. Multivariate analysis was again done by taking the top four significant variables obtained from the multivariate analysis performed earlier, and the results are shown in Table 3.

It is evident from Table 3 that all the four variables are highly significant with p-value less than 0.01. It can be inferred from Table 3 that the most significant factor influencing recovery is the severity of the disease (indicated by initial status). The most important factors identified by the various machine learning methods

are initial status, healthcare worker or not, age and pre-existing diseases.

Discussion

The results of the important variables identified by machine learning models such as SVM, J48, naives bayes and CTree are consistent with the earlier studies [18–20,25]. The J48 decision tree developed by Al-Turaiki et al. identified that pre-existing disease and whether the patient is a healthcare professional or not are the main factors influencing recovery [20]. The study conducted based on MERS infection in South Korea identified age and comorbidities as vital factors in determining recovery [25].

On analysis of the data it has been observed that there is only a slight difference between the death rate in patients exposed to camels and patients not exposed to camels. Hence the feature animal exposure turned out to be an insignificant attribute. However, it is to be specially noted that nearly 80% of the recovered patients who had exposure to camels were initially in a stable condition. This may have contributed immensely to recovery thereby making camel exposure as an insignificant factor. Nearly 10% of the MERS infected patients exhibited no symptoms and had recovered from the disease. It has been observed from the data obtained that all the asymptomatic patients were less severely infected and a majority of them acquired infection from others. The death rate among healthcare workers is negligible. This is due to increased infection awareness, early diagnosis and protective measures taken [26]. The univariate analysis of survival yields the result that gender is a significant factor and male are more likely to die. There is a considerable difference between the number of male and female infected patients. Furthermore the survival rate in women are more than men. This may be due to the following reasons:

i Nearly 33% of the infected women are healthcare personnel while it is only 8% in the case of men.
ii The social set up in the country makes men more prone to the disease than women. [27]
iii Compared to women, men are most likely to be in direct contact with camels. As per the data analysed nearly 94% of the patients in contact with camels were men.

However, it is evident from Table 2 that on performing multivariate analysis, gender feature turns out to be least significant [11]. This is due to the fact that attributes such as age, pre-existing diseases and critical initial status shield the gender attribute. It is obvious that irrespective of gender, the aforementioned characteristics adversely affect recovery. Univariate analysis shows that acquiring infection from others is a significant feature. The death rate due to primary infection is nearly double that of infection acquired from others. Both univariate and multivariate analysis shows that age is a significant attribute in determining recovery and older people are more likely to die. The mortality rate is around 50% in patients above 50 years of age. This is attributed to the fact that they are most likely to suffer from other comorbidities [12]. Both univariate and multivariate analysis reflect that pre-existing diseases have a negative impact on recovery. The above facts suggest that the elderly and people suffering from comorbidities demand special care in the treatment of MERS [18].

Initial status is the most significant attribute as per the results of both univariate and multivariate analysis. This result is consistent with the other machine learning model based methods used in the present work, the results of which are depicted in Fig. 1. Positive value for the estimate of initial status 'stable' indicates that the recovery rate is high in patients with less severe disease.

Conclusion

In this paper, the main factors influencing the recovery of MERS patients in Saudi Arabia are identified based on cases reported by Command and Control Centre of Ministry of Health, KSA. Machine learning models and statistical analysis are used for this purpose. The important factors identified include the patient's age, pre-existing diseases, initial status (severity of the infection) and whether he/she is a healthcare worker or not. Though MERS Co-V is zoonotic in nature, animal exposure is not a major factor leading to death. This is evidently due to stable initial condition of majority of the patients exposed to camels. The experiments conducted show that the severity of the disease is the most significant factor influencing recovery from the disease. It is apparent that the mortality rate can be reduced considerably, if the infected people seek medical attention at the earliest. Therefore widespread public awareness regarding the symptoms of MERS will play a vital role in increasing the recovery rate. Since the infection also spreads due to person to person contact, an effective monitoring system is highly essential to prevent the disease from spreading.

Funding

No funding sources.

Competing interests

None declared.

Ethical approval

Not required.

References

[1] Morens DM, Folkers GK, Fauci AS. The challenge of emerging and re-emerging infectious diseases. Nature 2004;430:242–9, http://dx.doi.org/10.1038/nature02759.
[2] Zaki AM, Van Boheemen S, Bestebroer TM, Osterhaus ADME, Fouchier RAM. Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia. N Engl J Med 2012;367:1814–20, http://dx.doi.org/10.1056/NEJMoa1211721.
[3] Arabi YM, Bouchama A, Luke T, Baillie JK, Al Omari A, Hajeer AH, et al. Middle East respiratory syndrome. N Engl J Med 2017;376:586–94.
[4] Cotten M, Watson SJ, Kellam P. Transmission and evolution of the Middle East respiratory syndrome coronavirus in Saudi Arabia: a descriptive genomic study. Lancet 2013;382:1993–2002.
[5] Gossner C, Danielson N, Gervelmeyer A, Berthe F, Faye B, Kaasik Aaslav K, et al. Human dromedary camel interactions and the risk of acquiring zoonotic Middle East respiratory syndrome coronavirus infection. Zoonoses Public Health 2016;63:1–9, http://dx.doi.org/10.1111/zph.12171.
[6] Azhar EI, El-Kafrawy SA, Farraj SA, Hassan AM, Al-Saeed MS, Hashem AM, et al. Evidence for camel-to-human transmission of MERS coronavirus. N Engl J Med 2014;370:2499–505, http://dx.doi.org/10.1056/NEJMoa1401505.
[7] Kasem S, Qasim I, Al-Hufofi A, Hashim O, Alkarar A, Abu-Obeida A, et al. Cross-sectional study of MERS-CoV-specific RNA and antibodies in animals that have had contact with MERS patients in Saudi Arabia. J Infect Public Health 2018;11:331–8, http://dx.doi.org/10.1016/j.jiph.2017.09.022.
[8] Assiri A, McGeer A, Perl TM, Price CS, Al Rabeeah AA, Cummings DAT, et al. Hospital outbreak of middle east respiratory syndrome coronavirus. N Engl J Med 2013;369:407–16, http://dx.doi.org/10.1056/NEJMoa1306742.
[9] Memish ZA, Zumla AI, Al-Hakeem RF, Al-Rabeeah AA, Stephens GM. Family cluster of middle east respiratory syndrome coronavirus infections. N Engl J Med 2013;368:2487–94, http://dx.doi.org/10.1056/NEJMoa1303729.
[10] Omrani AS, Matin MA, Haddad Q, Al-Nakhli D, Memish ZA, Albarrak AM. A family cluster of middle east respiratory syndrome coronavirus infections related to a likely unrecognized asymptomatic or mild case. Int J Infect Dis 2013;17:668–72, http://dx.doi.org/10.1016/j.ijid.2013.07.001.
[11] Memish ZA, Cotten M, Watson SJ, Kellam P, Zumla A, Alhakeem RF, et al. Community case clusters of Middle East respiratory syndrome coronavirus in Hafr Al-Batin, Kingdom of Saudi Arabia: a descriptive genomic study. Int J Infect Dis 2014;23:63–8, http://dx.doi.org/10.1016/j.ijid.2014.03.1372.
[12] Hui DS, Azhar EI, Kim Y-J, Memish ZA, Oh M, Zumla A. Middle east respiratory syndrome coronavirus: risk factors and determinants of primary, household, and nosocomial transmission. Lancet Infect Dis 2018;3099:1–11, http://dx.doi.org/10.1016/S1473-3099(18)30127-0.

[13] Breban R, Riou J, Fontanet A. Interhuman transmissibility of Middle East respiratory syndrome coronavirus: estimation of pandemic risk. Lancet 2013;382:694–9, http://dx.doi.org/10.1016/S0140-6736(13)61492-0.
[14] Al-Mohrej OA, Al-Shirian SD, Al-Otaibi SK, Tamim HM, Masuadi EM, Fakhoury HM. Is the Saudi Public aware of Middle East respiratory syndrome? J Infect Public Health 2016;9:259–66, http://dx.doi.org/10.1016/j.jiph.2015.10.003.
[15] Bawazir A, Al-Mazroo E, Jradi H, Ahmed A, Badri M. MERS-CoV infection: mind the public knowledge gap. J Infect Public Health 2018;11:89–93, http://dx.doi.org/10.1016/j.jiph.2017.05.003.
[16] Alsahaf AJ, Cheng AC. Knowledge, attitudes and behaviour of healthcare worker in the Kingdom of Saudi Arabia to MERS coronavirus and other emerging infectious diseases. Int J Environ Res 2016;13:1–8, http://dx.doi.org/10.3390/ijerph13121214.
[17] Yang Y-M, Hsu C-Y, Lai C-C, Yen M-F, Wikramaratna PS, Chen H-H, et al. Impact of comorbidity on fatality rate of patients with middle east respiratory syndrome. Sci Rep 2017;7:1–9, http://dx.doi.org/10.1038/s41598-017-10402-1.
[18] Rivers CM, Majumder MS, Lofgren ET. Risks of death and severe disease in patients with Middle East respiratory syndrome coronavirus, 2012–2015. Am J Epidemiol 2016;184:460–4, http://dx.doi.org/10.1093/aje/kww013.
[19] Ahmed AE. The predictors of 3- and 30-day mortality in 660 MERS-CoV patients. BMC Infect Dis 2017;17:1–8, http://dx.doi.org/10.1186/s12879-017-2712-2.
[20] Al-Turaiki I, Alshahrani M, Almutairi T. Building predictive models for MERS-CoV infections using data mining techniques. J Infect Public Health 2016;9:744–8, http://dx.doi.org/10.1016/j.jiph.2016.09.007.
[21] Abdullah M, Altheyab MS, Lattas AM, Algashmari WF. MERS-CoV disease estimation (MED) a study to estimate a MERS-CoV by classification algorithms. In: de Alencar MS, editor. Commun. manag. Inf. Technol. 1st ed. Taylor and Francis; 2016. p. 633–8.
[22] Li EY, Tung CY, Chang SH. The wisdom of crowds in action: forecasting epidemic diseases with a web-based prediction market system. Int J Med Inform 2016;92:35–43, http://dx.doi.org/10.1016/j.ijmedinf.2016.04.014.
[23] Wiemken TL, Furmanek SP, Mattingly WA, Guinn BE. Predicting 30-day mortality in hospitalized patients with community-acquired pneumonia using statistical and machine learning approaches predicting 30-Day mortality in hospitalized patients with community. Univ Louisv J Respir Infect 2017;1:50–6, http://dx.doi.org/10.18297/jri/vol1/iss3/10/Available.
[24] Quinlan R. C4.5: programs for machine learning. San Mateo, CA: Morgan Kaufmann Publishers; 1993.
[25] Majumder MS, Kluberg SA, Mekaru SR, Brownstein JS. Mortality risk factors for middle east respiratory syndrome outbreak, South Korea, 2015. Emerg Infect Dis 2015;(21):2088–90, http://dx.doi.org/10.3201/eid2111.151231.
[26] Ahmed AE. Diagnostic delays in 537 symptomatic cases of MERS-CoV infection in Saudi Arabia. Int J Infect Dis 2017;62:47–51, http://dx.doi.org/10.1016/j.ijid.2017.07.00.
[27] Chen X, Chughtai AA, Dyda A, Macintyre CR. Comparative epidemiology of Middle East respiratory syndrome coronavirus (MERS-CoV) in Saudi Arabia and South Korea. Emerg Microbes Infect 2017;6:1–6, http://dx.doi.org/10.1038/emi.2017.40.