

Age Structured Mixture Model for Early COVID-19 Spread: A Zimbabwean Risk Factor Analysis

Chipo Zidana ^{1*}, Masilin Gudoshava ², Sarudzai Portia Showa ²

¹Botswana International University of Science and Technology, BOTSWANA

²National University of Science and Technology, ZIMBABWE

*Corresponding Author: zidanac@biust.ac.bw

Citation: Zidana C, Gudoshava M, Showa SP. Age Structured Mixture Model for Early COVID-19 Spread: A Zimbabwean Risk Factor Analysis. Journal of Contemporary Studies in Epidemiology and Public Health. 2020;1(1):ep20003. <https://doi.org/10.30935/jconseph/8442>

ABSTRACT

Unique severe acute respiratory syndrome Coronavirus 2 (SARS-CoV-2/COVID-19) prevention measures to distinct age, geographical and community groupings can only be effectively and efficiently implemented with a clear understanding on dynamics of the disease. Dynamics include disease spread, different risk factors and their level of influence and individual attributes that aid the spread. The paper aims at determining the major COVID-19 spread risk factors in Zimbabwe by identifying individual, age and community groupings, their risk levels given the complex heterogeneous population. COVID-19 data for 37 individuals as provided by the Ministry of Health and Child Care (MoHCC) for the period from 20 March - 14 May 2020 is used. Generalised Mixture Models were implemented to achieve the objectives. Results show that gender, age, mode of infection and history of travel were the main predictors of COVID-19 spread in Zimbabwe. However, their effects were distributed differently across two clusters. Children (0-14) years, females and those with imported infections were among high level risk spread groups. Whilst low risk groups consist non travelers, males and those infected by local transmission. We thus recommend that the Zimbabwean government need to prioritise children, females, and non-travelers when implementing prevention measures.

Keywords: COVID-19 spread, age structured, spread risk factors, finite mixture models, COVID-19 prevention

Received: 13 Jun. 2020 ♦ Accepted: 13 Jul. 2020

INTRODUCTION

As Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) continues to be a problem world over, it is paramount to understand the major determinants of the spread of the virus in Zimbabwe. To date Corona virus disease 2019 (COVID-19) remains with no cure or vaccine. The COVID-19 pandemic has resulted in so many fears, myths, misrepresentations and misconceptions [1, 2]. Different governments and authorities have been putting different measures to help limit the spread of the disease [3, 4]. However, enough knowledge needs to be obtained for the different measures to be effective. Adequate effective measures will require capturing the complexity of COVID-19 at many levels including individual level attributes, community level attributes together with their interactions as risk factors based on unique economies, and environments [5].

There have been different schools and myths on the dynamics of COVID-19 in children, their risk of spreading the disease and how to keep them healthy in COVID-19 times [5, 6]. We intend to assess the risk of spreading COVID-19 by children as well as all other different age categories in Zimbabwe. Although COVID-19 deterministic age structured models have been developed in several countries in order to inform on the implementation of isolation strategies per age group [7,

8], same control measures in different countries have resulted in different courses of the disease dynamics and contrasting impacts as alluded by [9]. Understanding the age structure risks peculiar to Zimbabwe is therefore of paramount importance to enhance the Zimbabwe COVID-19 prevention strategies. We divided the age variable into five (5) different categories to capture the heterogeneous age structures in Zimbabwe. These categories are the same as those used by the Ministry of Health and Child Care (MoHCC) in Zimbabwe. This is done to achieve the objective and determine associated risk for each age category. Zimbabwe is divided into different regions which offer different lifestyles, varying hygienic standards and health facilities. These regions include rural, urban, peri urban, farms, mines etc. Therefore, efficient and effective measures for Zimbabwe can only be implemented if there is an understanding of the COVID-19 dynamics and risks in these different regions.

It is our aim therefore to demystify the fears and misconception in Zimbabwe by determining the major predictors of COVID-19 by age and region using the data availed by the Ministry of health and Child Care daily reports on www.mohcc.gov.zw/ and their corresponding social media platforms (<https://twitter.com/MoHCCZim>). Our objectives are to

- Determine COVID-19 major predictors in Zimbabwe,

Table 1. Variable Description

Variable	Abbreviation	Description
Number of Active cases per diagnosis date	DTcases	The total number of active cases in Zimbabwe as per day of diagnosis
Case number	Case	Case Identification number
Date of Diagnosis	DD	Date on which the diagnosis was made
Number of Tests	NT	Total number of test (PCR and RRT) conducted per diagnosis day
Gender	Sex	0 Female 1 Male
History of Travel	travel	History of travel from December 2019 to date of diagnosis 0 No 1 Yes
Age	A_cat	The individual age category 1. 0-14 years 2. 15-29 years 3. 30-44 years 4. 45-59 years 5. >60 years
Mode of Infection	INF_cat	The method by which an individual was infected 0 Unknown 1 Imported 2 Local transmission
Location	Location	Residential city/district
Province	P_loc	Provincial location Bulawayo (BYO) Harare (HRE) Mashonaland East (MSE) Mashonaland west (MSE) Matabeleland North (MTN)

- Identify the different risk group of COVID-19, and hence identify with higher risk and lower risk heterogeneous populations in Zimbabwe.

MATERIALS AND METHODS

Data

The work considered thirty-seven (37) individual data profiles as provided by the MoHCC in Zimbabwe recorded in the period from 20 March to 14 May 2020. Individual profiles included eleven (11) variables as given by **Table 1**. The number of active cases per day of diagnosis was used as the dependent variable to access how it is affected by the predictor/risk variables. COVID-19 associated risks cannot be treated as a one blanket suit all scenario. To assess the effects on children, the age variable was further categorised using five (5) age layers provided by MoHCC in Zimbabwe as shown in **Table 1**.

Although all the 10 provinces in Zimbabwe were considered for capturing the number of tests conducted, only provinces with active cases will be used as we are using number of active cases per diagnosis date. The total number of tests per day considered consist of combined Polymerase Chain Reaction (PCR) and Rapid Results Test done per day. History of travel captures whether an active individual once travelled outside Zimbabwe since December 2019 or not, whilst mode of transmission has three categories: either imported due to travel, locally transmitted as contact case or unknown if there is no evidence of the first two scenarios. Location refers to the residential region someone resides in and the province thereof.

Summary by age and province

The study group had more of middle-aged people in categories 2 and 3, few elderly people and only 2 children (0-14 years). The median

number of cases for the children was 5, whilst for the over 60 years was 3. Active cases in Zimbabwe are distributed within 5 provinces for the referred period. Of the 5 provinces, Harare has the highest number of cases standing at 14, followed by Bulawayo (12), Mashonaland East (6), Mashonaland West (4) and lastly Matabeleland North recording only a single case. The highest median number of cases was 3 in Bulawayo and Mashonaland cases with Mashonaland east having the median number of 1. The descriptive statistics on the number of COVID-19 active cases in Zimbabwe by location and Age category is shown in **Table 2**.

Statistical Analysis Models

In this study, risk associated with increasing the number of active cases (spread) by each of the predictor variables on the COVID-19 is implemented via Generalized linear Mixture models (GLM Mixture) models. We intend to explore the effects of Age and location differently to come up with a holistic understanding of the age structure and location on the spread in COVID-19 in Zimbabwe. GLM Mixture models allows us to measure risk factors by considering heterogeneous risk groups comprising of similar individual attributes as in [10]. The groupings will also enable us to infer into level of risk (high, medium or low) based on their individual composition. All the statistical analysis was done in R version 3.6.9 at 5% level of significance using flexmix packages [11] for the GLM Mixture models.

GLM mixture model

A GLM Mixture Regression model is used in order to identify the risk groups, individual level risk, community/location level risk and age level risk effects of each predictor and level of risk associated. This is due to its high ability to capture heterogenous attributes without having to give a lot of sometimes unrealistic assumptions on the data. Mixture models also works better on diseases with complex diagnosis and

Table 2. Data Summary

Category	Age (years)				Province				
	count	sd_size	median	IQR	Province	count	sd_size	median	IQR
1 (0-14)	2	0.00	5	0	BYO	12	1.61	3.0	3.00
2 (15-29)	11	1.18	2	1	HRE	14	0.83	1.5	1.00
3 (30-44)	10	1.32	2	2	MSE	6	0.52	1.0	0.75
4 (45-59)	9	1.45	1	2	MSW	4	1.00	3.0	0.50
5 (> 60)	5	0.89	3	1	MTN	1	NaN	1.0	0.00

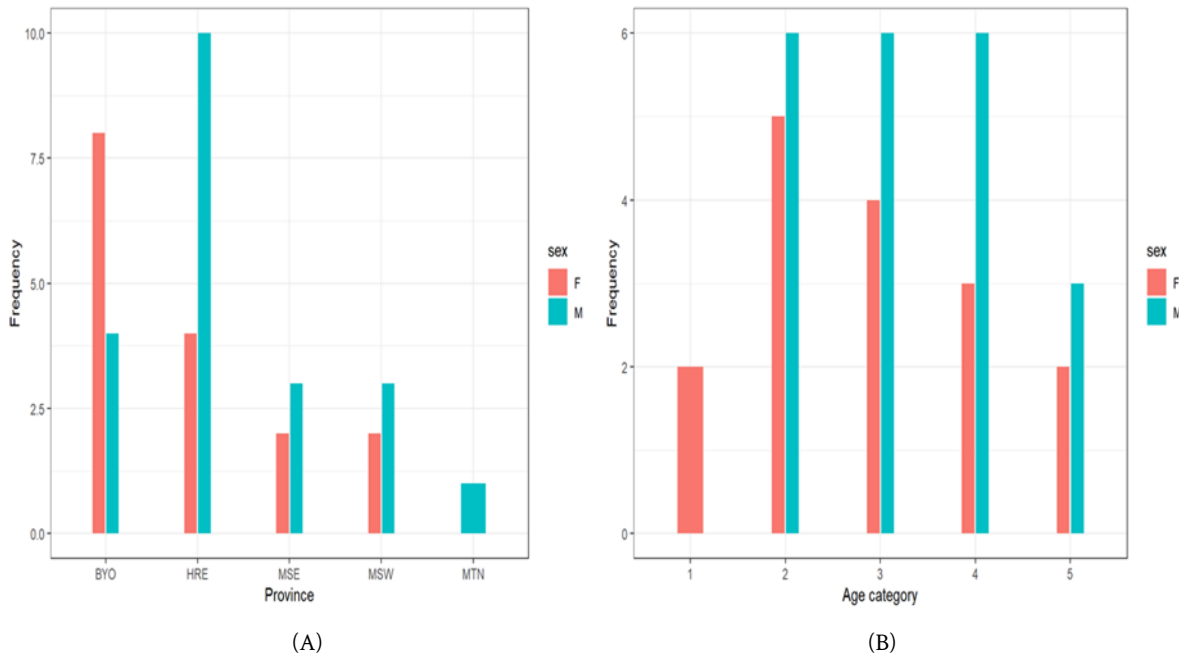


Figure 1. Distribution of active cases by province, age category and gender

circumstances as evidenced in COVID-19. The GLM Mixture model can be represented by the model

$$h(y|x, \Phi) = \sum_{k=1}^K \pi_k f(y|x, \theta_k) \tag{1}$$

such that we have

$$\pi_k \geq 0, \sum_{k=1}^K \pi_k = 1$$

where y represent the response variable in this case, number of active COVID-19 cases diagnosed per day with conditional mixture density h, x is a vector of risk predictor variables, π_k is the prior probability of an individual being in component k , whilst θ_k represent the component specific parameter vector with a density distribution f and finally $\Phi = (\pi_1, \dots, \pi_k, \theta_1, \dots, \theta_k)$ is a vector containing all parameters. The parameter estimates will be done using the Expectation Maximisation (EM) algorithm.

ANALYSIS AND RESULTS

The number of active cases by provincial location and age categories show that indeed there are differences on how age groups are related to the number of counts per diagnosis (Figure 1). Figure 1A shows that more females were infected in Bulawayo than any other province whilst more males were infected in all the other provinces with the highest being in Harare. Figure 1B shows that more males

were infected across all the age categories except category 1 that consists of children (0-14 years) where only females were infected.

GLM Mixture Model Results

The GLM mixture model enables us to not only identify the main COVID-19 predictors but to capture the complexity of the individual and group level heterogeneous characteristics. In this case individual level characteristics across age groups and community level of risk of spread could be identified. A two-component risk model based on individual characteristics was identified using both Poisson and Gaussian link function. Results showed that a Gaussian Mixture model with two components was more appropriate due to its low AIC value. The Gaussian GLM Mixture model had a lower AIC value of 65.75 compared to the Poisson GLM Mixture (AIC = 163.76). Gaussian GLM Mixture model results indicated that the two clusters were well separated with component ratio for component 1 being (1.00) a scenario clearly shown by Figure 2A. The fact that the rootogram in Figure 2A had its highest peaks at the ends for both components is a good indication that cluster/components were well separated. Overall, component 1 had 26 observations with a probability of $\pi_1 = 0.628$ whilst component 2 had 11 observations and probability of 0.372. The variability of component 1 was much higher than that of component 2 as indicated by the standard deviation of 2.278 and 0.211 respectively. Results in Table 3 shows that the total number of tests does not affect the spread of COVID-19, so we removed the factors from the analysis. Parameter estimates for component 1 in Table 3 showed the following: 1) children were significantly more likely to spread COVID-19 by 2.13,

Table 3. Parameter Estimates for Mixture Model

Component 1: $\pi_1 = 0.628, \sigma_1 = 0.278, n = 26$ ratio = 1				
	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	5.202318	0.179321	29.0112	< 2.2e-16 ***
A_cat2	-2.134268	0.174088	-12.2597	< 2.2e-16 ***
A_cat3	-2.152280	0.175161	-12.2875	< 2.2e-16 ***
A_cat4	-2.329225	0.176040	-13.2313	< 2.2e-16 ***
A_cat5	-1.026736	0.196124	-5.2351	1.649e-07 ***
sexM	-1.115381	0.080537	-13.8493	< 2.2e-16 ***
travel1	-2.015194	0.110609	-18.2190	< 2.2e-16 ***
INF_cat1	0.901114	0.118295	7.6175	2.586e-14 ***
INF_cat2	-0.202318	0.112850	-1.7928	0.07301
P_locHRE	0.183850	0.124089	1.4816	0.13845
P_locMSE	-0.811440	0.111951	-7.2482	4.224e-13 ***
P_locMSW	-0.954011	0.120960	-7.8870	3.095e-15 ***
P_locMTN	0.179423	0.255425	0.7024	0.48240

Component 2: $\pi_1 = 0.372, \sigma_1 = 0.211, n = 11$, ratio = 0.55				
	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	3.9849e+00	2.3769e-01	16.7651	< 2.2e-16 ***
A_cat2	-2.7301e-01	2.0196e-01	-1.3518	0.1764
A_cat3	-1.0937e-07	2.1998e-01	0.0000	1.0000
A_cat4	2.7301e-01	2.0196e-01	1.3518	0.1764
A_cat5	-1.9913e+00	2.1963e-01	-9.0665	< 2.2e-16 ***
sexM	1.1655e+00	1.5987e-01	7.2903	3.093e-13 ***
travel1	-2.1814e+00	2.1488e-01	-10.1516	< 2.2e-16 ***
INF_cat1	2.7057e+00	2.1596e-01	12.5290	< 2.2e-16 ***
INF_cat2	1.0151e+00	1.7081e-01	5.9428	2.801e-09 ***
P_locHRE	-3.2504e+00	1.8284e-01	-17.7771	< 2.2e-16 ***
P_locMSE	-4.9316e+00	2.2350e-01	-22.0654	< 2.2e-16 ***
P_locMSW	-4.1478e+00	3.4435e-01	-12.0455	< 2.2e-16 ***
P_locMTN	-4.6748e+00	3.2561e-01	-14.3568	< 2.2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

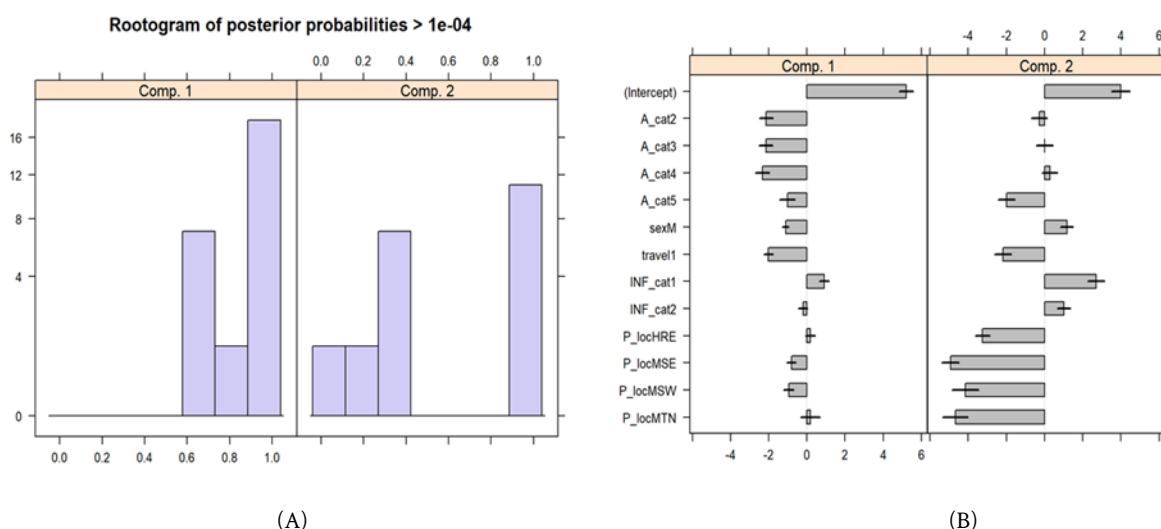


Figure 2. GLM Mixture Model: A) Cluster Rootogram B) Parameter Estimates Confidence intervals

2.15, 2.32 and 1.03 times compared to the 15-29, 30-44, 45-59 and over 60-year categories, respectively, 2) males are significantly 1.12 times less likely to spread COVID-19 compared to females, 3) Travelers in component 1 are 2.02 less likely to spread COVID-19 than non travelers, 4) Those who had imported infections are 0.90 times

significantly more likely to spread COVID-19 than those whose mode of infection is unknown, 5) those infected via local transmission were 0.20 times less likely to spread the disease compared to those whose mode of infection was unknown although the difference is insignificant, 6) Bulawayo residents in component 1 were more likely

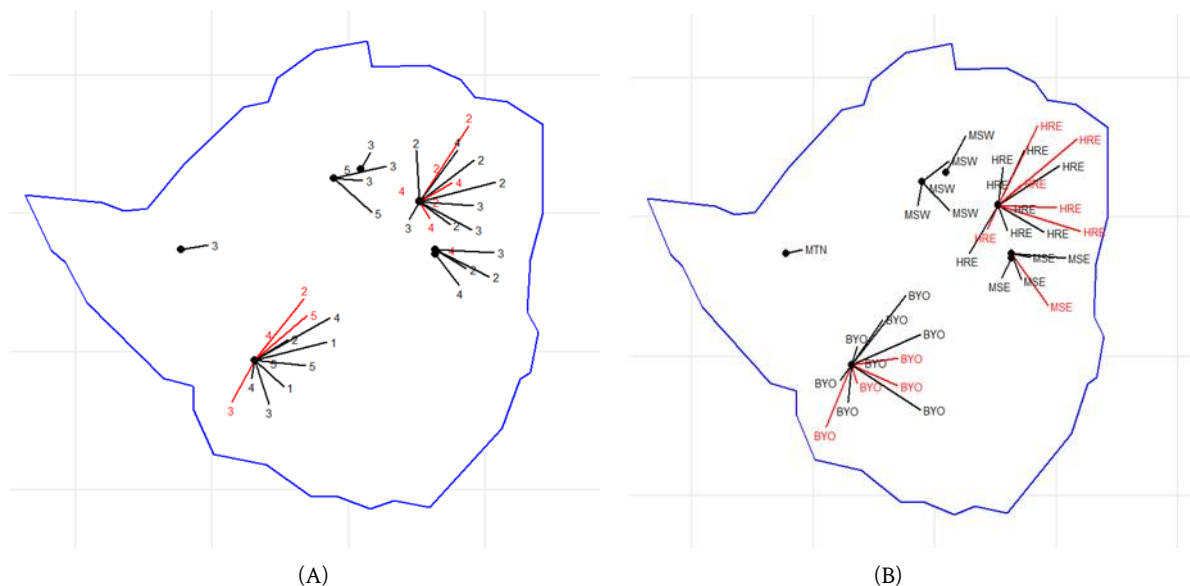


Figure 3. Cluster distribution by age category and provincial location

to increase the number of COVID-19 active cases by 0.81 and 0.95 than those in Mashonaland East and Mashonaland West, respectively, 7) Although Harare and Matabeleland North residents were more likely to spread the disease than Bulawayo residents by 0.18 and 0.17 times, respectively, the differences were insignificant. Observing the individual characteristics of those which belongs to component 1 we can conclude that: the high-risk groups in this cluster consists of children, females, non-travelers and those who had imported infections. These were for individuals mainly from Harare, Mashonaland west and Matabeleland North provinces.

Individuals from component 2 showed the following: 1) Although children were more likely to spread the disease than 15-29, 30-44, and over 60-year categories, only the 60-year category was 1.99 times significant. Children were however 0.27 times less likely to spread COVID-19 than the 45-59 year category although this was insignificant, 2) Males in component 2 were 1.17 times significantly more likely to increase the number of active cases than females, 3) Travelers were still 2.18 less likely to spread the disease than non-travelers, 4) Those who had imported infections were and infected locally were 2.70 and 1.01 significantly more likely to spread COVID-19 than those with unknown transmission, 5) Bulawayo residents were 3.25, 4.93, 4.15 and 4.67 more likely to increase the COVID-19 cases than Harare, Mashonaland East, Mashonaland West and Matabeleland North provinces respectively. We can therefore conclude that the level of risk in component 2 is a bit lower than component 1 based on the magnitude of the parameter values. Secondly, inferring into individual attributes for those in component 2, we observed that high potential risk to spread COVID-19 consists of: children and elderly, males, non-travelers, those who had imported infections and those who got infected through local transmission. These were main characteristics for Bulawayo residents.

The confidence interval for model parameter for both components are shown by **Figure 2B**. Whilst age differences mainly characterize individuals in component 1, location differences mainly characterize individuals in component 2. We observed that component 1 parameter estimates were more positive than those in component 2 thus we can conclude that component 1 individuals pose more risk to spread

COVID-19 than those in component 2. Although in general children and non-travelers are more likely to spread the disease in both components, the high-risk cluster was uniquely associated with children, females and those with imported infections. We can overall associate high risk with Harare, Mashonaland West and Matabeleland North residents. Bulawayo and Mashonaland East residents can be categorized under the low risk cluster. This is a clear indication that in Zimbabwe, effective measures may have to give priority to children, gender and make sure that non travelers are protected from the spread. Low risk cluster, however, characterized by mainly Bulawayo and Mashonaland East residents where spread was mostly likely to be from males, elderly and those with either with imported or locally transmitted infections. The distribution of age groups by either high risk or low risk group is shown by **Figure 3A** whilst distribution by province is given by **Figure 3B**. It is evident from **Figure 3A** that children (represented by 1) are in cluster 2 which is the high risk cluster and has only imported cases mainly in Mashonaland West and Matabeleland provinces as indicated by **Figure 3B**. Low risk which is characterized by local transmissions consist mainly of Bulawayo and Mashonaland East residents.

DISCUSSION

A GLM Gaussian Mixture model was fitted to Zimbabwean COVID-19 data for the period from 20 March- 14 May 2020 as provided by the Ministry of Health and Child Care in Zimbabwe. The primary goal was to determine the major risk factors associated with the spread of COVID-19 given the heterogeneous age structure and locations found in Zimbabwe. The model was fitted to 37 individual data and the following 10 variables were considered: number of cases per day, total number of tests conducted per day, gender, age, history of travel, mode of infection, location, province and date of diagnosis. A mixture model was preferred due to its flexibility in handling complex heterogeneous problems. This model enabled us to identify different risk groups and their associated levels of risk. Age structure models were considered so that preventative measures will be better implemented on more risk age groups and province/locations

compared to just general one blanket fit all measures. Specific interest in this model was the risk of children in spreading the disease.

Results from the Gaussian mixture model classified individuals into two (2) groups based on their individual characteristics and hence risk levels of spreading the disease. We termed cluster 1 high risk cluster and cluster 2, the low-risk cluster, respectively. Whilst the major risk factors remain the same (gender, history of travel, mode of infection, province and age category) across clusters, their risk contribution was distributed differently depending on whether an individual is in the high risk or low risk cluster. Overall, the model showed the risk group predictors as being a female, a non-traveler, child, local infections and imported cases. The probability of getting into a high-risk cluster was (0.628), a much higher than the low risk cluster (0.372) an indication that in Zimbabwe COVID-19 is 0.256 more likely to be spread than controlled. Considering cluster 1 attributes we noticed that all age categories were likely to spread the disease although children a much higher potential to spread COVID-19 than any other age group. Females and those with imported infected were also among the high-risk groups in cluster 1 an observation with most Harare residents (which had the highest number of active cases), Mashonaland West and Matabeleland North provinces. We observed as alluded by [5] a mixture model distinguishes between individual level, community level, and group level risks associated by each individual in the spread of COVID-19. Whilst it is generally believed that men are at more risk for worse outcomes and deaths given the same prevalence with women [12], our results showed that in Zimbabwe women tend to spread COVID-19 more than men even when age is also being considered as a very important factor in the spread. It is evident therefore that in Zimbabwe children (0-14 years), those with imported infections and females have a higher risk of spreading COVID-19 disease. Based on our findings, we can therefore conclude the age structure population is important in understanding COVID-19 dynamics as alluded by [13,14]. In Zimbabwe, major prevention measures on the spread should also target children and females and imported infections management. It is also interesting to note that in Zimbabwe, the major risk of COVID-19 spread is by those infected outside the country are concentrated in the capital city Harare.

Thus, government may need to either keep the borders closed to avoid imported infection and in cases where it is unavoidable, travelers entering Zimbabwe must be severely quarantined and monitored. Measures also need to be implemented targeting different gender groups as the model predicted that in Zimbabwe, females are more likely to spread COVID-19 than their male counterparts. The low risk cluster however, consisted mainly of individuals from Bulawayo and Mashonaland East, males, non-travelers and those who had local transmissions. Again, measures that minimize local transmission may be implemented like isolation centers to cater for those infected until they heal. The differences in the provinces although exhibited as risk to a lesser extent should also be explored. Considering the heterogeneous difference of Zimbabwe's residential set up, this could have been attributed by the differences in sanitary conditions in the different areas. Since COVID-19 spread is highly related to hygienic conditions, improvement in hygiene in residential predicted to have a high risk may curb the spread of the infection.

CONCLUSION

COVID-19 in Zimbabwe has been largely due to imported cases and lesser extend local transmissions. High risk groups for the spread of the disease include, children, women non-travelers. Thus, therefore these social groupings should be thoroughly considered when authorities are to come up with any meaningful prevention measures. Overall, the difference in the residential locations although they contribute to spread, they pose a lesser risk to spread of COVID-19 compared to age differences.

REFERENCES

1. Roy S. Low-income countries are more immune to COVID-19: A misconception. *Indian J Med Sci.* 2020 Apr 30;72(1):5-7. (doi: 10.25259/IJMS_26_2020).
2. Mamun MA, Griffiths MD. First COVID-19 suicide case in Bangladesh due to fear of COVID-19 and xenophobia: Possible suicide prevention strategies. *Asian J Psychiatr.* 2020 Jun 1;51:102073. (doi: 10.1016/j.ajp.2020.102073).
3. Hale T, Angrist N, Kira B, Petherick A, Phillips T, Webster S. Variation in government responses to COVID-19 [Internet]. 2020 [cited 2020 Jun 3]. Available at: www.bsg.ox.ac.uk/covidtracker
4. WHO. Coronavirus disease (COVID-19) [Internet]. [cited 2020 Jun 3]. Available at: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200601-covid-19-sitrep-133.pdf?sfvrsn=9a56f2ac_4
5. Baral SD, Mishra S, Diouf D, Phanuphak N, Dowdy D. The public health response to COVID-19: balancing precaution and unintended consequences. *Annals of Epidemiology.* 2020;46:12-3. (doi: 10.1016/j.annepidem.2020.05.001).
6. Fore HH. A wake-up call: COVID-19 and its impact on children's health and wellbeing. *Lancet Glob Heal.* 2020 May. (doi: 10.1016/S2214-109X(20)30238-2).
7. Robertson T, Carter ED, Chou VB, Stegmuller AR, Jackson BD, Tam Y, et al. Early estimates of the indirect effects of the COVID-19 pandemic on maternal and child mortality in low-income and middle-income countries: a modelling study. *Lancet Glob Heal.* 2020 May 12. (doi: 10.1016/S2214-109X(20)30229-1).
8. Prem K, Liu Y, Russell TW, Kucharski AJ, Eggo RM, Davies N, et al. The effect of control strategies to reduce social mixing on outcomes of the COVID-19 epidemic in Wuhan, China: a modelling study. *Lancet Public Heal.* 2020 May 1;5(5):e261-70. (doi: 10.1101/2020.03.09.20033050).

9. Canabarro A, Tenorio E, Martins R, Martins L, Brito S, Chaves R. Data-Driven Study of the the COVID-19 Pandemic via Age-Structured Modelling and Prediction of the Health System Failure in Brazil amid Diverse Intervention Strategies. medRxiv [Internet]. 2020 Apr 15 [cited 2020 Jun 3];2020.04.03.20052498. Available at: <http://medrxiv.org/content/early/2020/04/08/2020.04.03.20052498.abstract>
10. Mufudza C, Erol H. Poisson Mixture Regression Models for Heart Disease Prediction. *Comput Math Methods Med*. 2016;4083089:10. (doi: 10.1155/2016/4083089).
11. Grün B, Leisch F. FlexMix: An R package for finite mixture modelling. *R News* [Internet]. 2007;7(1):8-13. Available at: <http://cran.r-project.org/doc/Rnews/>
12. Jin JM, Bai P, He W, Wu F, Liu XF, Han DM, et al. Gender Differences in Patients With COVID-19: Focus on Severity and Mortality. *Front Public Heal*. 2020 Apr 29;8. (doi: 10.3389/fpubh.2020.00152).
13. Dudel C, Riffe T, Acosta E, van Raalte AA, Strozza C, Myrskylä M. Monitoring trends and differences in COVID-19 case fatality rates using decomposition methods: Contributions of age structure and age-specific fatality. (doi: 10.46610/JoTS.2020.v05i03.003).
14. Singh R, Adhikari R. Age-structured impact of social distancing on the COVID-19 epidemic in India [Internet]. Available at: <https://github.com/rajeshrinet/pyross>