COVID-19 infection wave mortality from surveillance data in the Philippines using machine learning

Short Title: COVID-19 infection wave mortality in the Philippines

Authors:
Julius R Migriño, Jr[a,b], Ani Regina U Batangan[a] and Rizal Michael R Abello[a]

[a] San Beda University College of Medicine, Manila, Philippines
[b] Ateneo de Manila University School of Medicine and Public Health, Pasig, Philippines

Correspondence to: Julius R. Migriño, Jr. (email: jrmjrmd-1@yahoo.com)

ABSTRACT

Objective: The Philippines has had several COVID-19 infection waves brought about by different strains and variants of SARS-CoV-2. This study aimed to describe COVID-19 outcomes by infection waves using machine learning.

Methods: We used a cross-sectional surveillance data review design using the DOH COVID DataDrop data set as of September 24, 2022. We divided the data set into infection wave data sets based on the predominant COVID-19 variant(s) of concern during the identified time intervals: ancestral strain (A0), Alpha/Beta variant (AB), Delta variant (D), and Omicron variant (O). Descriptive statistics and machine learning models were generated from each infection wave data set.

Results: Our final data set consisted of 3 896 206 cases and ten attributes including one label attribute. Overall, 98.39% of cases recovered while 1.61% died. The Delta wave reported the most deaths (43.52%), while the Omicron wave reported the least (10.36%). The highest CFR was observed during the ancestral wave (2.49%), while the lowest was seen during the Omicron wave (0.61%). Higher age groups generally had higher CFRs across all infection waves. The A0, AB and D models had up to four levels with two or three splits for each node. The O model had eight levels, with up to 16 splits in some nodes. Of the ten attributes, only age was included in all the decision tree models, while region of residence was included in the O model. F-score and specificity were highest using naïve Bayes in all four data sets. Area under the curve (AUC) was highest in the naïve Bayes models for the A0, AB and D models, while sensitivity was highest in the decision tree models for the A0, AB and O models.

Discussion: The ancestral, Alpha/Beta and Delta variants seem to have similar transmission and mortality profiles. The Omicron variant caused lesser deaths despite being more transmissible. Age remained a significant predictor of death regardless of infection wave. We recommend constant timely analysis of available data especially during public health events and emergencies.

Keywords: *COVID-19, mortality, variant, machine learning, surveillance*

INTRODUCTION

The Philippines has been considered a hotspot for the coronavirus disease 2019 (COVID-19) in the Western Pacific region.[1] As of December 1, 2022, the country's Department of Health (DOH) has reported a total of 4,037,547 cases, including 64,658 reported deaths.[2] Meanwhile, the World Health Organization (WHO) has tallied 639,572,819 confirmed cases and 6,615,258 deaths globally.[3] The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is the primary etiologic agent of COVID-19 infection. The infection causes symptoms like cough, colds, fever, dyspnea and dysgeusia, and may progress to be more life-threatening if it presents with complications such as shock and organ failure. COVID-19 mortality is influenced by several factors like advanced age, sex, presence of pre-existing comorbid illness, and history of smoking and alcohol consumption.[1] More recently, studies have surfaced highlighting the differences in mortality rates among cases with different vaccination statuses[4,5] and among those who were previously infected.[5] According to SeyedAlighani and his colleagues (2021), mortality rates are additionally influenced by adequacy of health care delivery, political decisions, and epidemiological characteristics of the affected population.[6]

Generally, viruses evolve to become more transmissible, regardless of severity.[7] The ancestral strain was the original SARS-CoV-2 virus which originated in China. The virus has been persistent in its infection rates due to its intrinsic capability to replicate and mutate. These spontaneous mutations are products of viral RNA replication errors within the host cell resulting in the appearance of multiple variants.[8] As of December 2022, there have been five recognized circulating SARS-CoV-2 variants of concern (VOCs): Alpha, Beta, Gamma, Delta and Omicron. These VOCs have appeared in infection waves among different countries in varying timelines, but their designation as VOCs were December 2020 (Alpha and Beta), January 2021 (Gamma), May 2021 (Delta) and November 2021 (Omicron).[9] Recent studies have characterized the different VOCs in terms of their transmissibility and severity. For instance, while the Delta variant evolved to become more transmissible than previous variants, several studies report similar hospitalization and mortality rates among the different infection waves.[10–12] The Omicron variant, on the other hand, proved to be even more highly transmissible compared to the previous variants, but has shown the lowest hospitalization and mortality rates.[13] The observed differences in transmission and severity among COVID-19 variants is possibly related to the increased immunity among the people infected, either through vaccination or previous infection waves.[7]

As of October 8, 2022, there had been a total of 22,400 SARS-CoV-2 sequences shared by the Philippines in the Global Initiative on Sharing All Influenza Data (GISAID) COVID-19 sequence repository, which accounts for 0.57% of all cases.[14] Tracking of relative frequencies of variants from sequenced COVID-19 cases show estimated time frames of the upsurge of specific variants: the ancestral strain was predominant (i.e., made up more than 50% of all sequenced samples) until about February 2021; the Alpha and Beta variants were concurrently predominant starting March until June 2021; the Delta variant was predominant from July until November 2021; starting from December 2021 until present, the Omicron variant and its subvariants were predominant.[15]

Machine learning is often used for health in the analysis of large datasets and the prediction of outcomes based on a variety of inputs. Such applications of machine learning include identification of disease from clinical symptoms or laboratory results, as well as in treatment of diseases and facilitation of administrative processes. Such techniques have been used to aid in treatment of cancer, pneumonia, diabetes and other diseases, including COVID-19, wherein they can give more than 90% accuracy in prediction and forecasting.[16]

Early prediction of COVID-19 mortality risks may help mitigate the effect of the pandemic by providing evidence for efficient resource allocation and proper patient treatment plans,[17] and has been the topic of several researches.[17–19] Most studies relied on medical records from admitted patients, relying on demographic, clinical and laboratory features to generate predictive models for patient prognosis. Some examples of machine learning algorithms used in COVID-19 research include logistic regression,[18] support vector machines,[17] and decision tree ensembles (e.g., CatBoost, XGBoost, Random Forest).[19] A previous study utilized a publicly available national surveillance dataset to predict COVID-19 mortality in the Philippines and identified age and history of hospital admission as significant predictors of disease outcome, but the study was limited to the early part of the pandemic wherein only the ancestral strain

was present in the population.[20] This study aimed to describe COVID-19 outcomes by infection waves using machine learning.

METHODS

The study utilized a cross-sectional, documents review design. Data from the publicly available DOH COVID Data Drop database for September 24, 2022[2] was utilized in this study. The database represented all reported COVID-19 cases by reverse transcription polymerase chain reaction of respiratory swabs and was updated daily by the DOH Epidemiology Bureau. The raw data set contained 3 934 777 cases and 22 attributes. Exploratory analysis was performed to screen cases and attributes in the raw data set. Ten attributes were included in the model generation which included *Age*, *Sex*, *Admitted*, *RegionRes*, *ProvRes*, *CityMunRes*, *BarangayRes*, *Quarantined*, *Pregnanttab* and *RemovalType*. Another attribute *Age_Group* was generated to reclassify *Age* into nine bins based on the US CDC classification for descriptive statistics. Another attribute, *DateRepConf*, was retained only for splitting of the data sets (below). Missing values for *Pregnanttab* were recoded as "(N/A)" for cases with *Sex*=MALE. Missing values for *BarangayRes*, *CityMunRes* and *ProvRes* were recoded as "ROF" for all cases where *RegionRes*="ROF". Cases with missing values for *Age* and *RemovalType* were dropped from the data set to generate the final data sets. Details of the exploratory analysis can be found in **Supplementary Information A**.

The final dataset was split into four (4) data sets according to *DateRepConf*, where each data set represents the predominant COVID-19 variant(s) of concern during those time intervals as reported by GISAID[15]. The details of the four data sets are listed below:

1. A0 data set (predominant strain: ancestral; start date: January 30, 2020; end date: February 28, 2021);
2. AB data set (predominant variants: Alpha and Beta; start date: March 1, 2021; end date: June 30, 2021);
3. D data set (predominant variant: Delta; start date: July 1, 2021; end date: November 30, 2021);
4. O data set (predominant variant: Omicron; start date: December 1, 2021; end date: September 24, 2022)

Descriptive statistics such as means, standard deviations, frequencies, case fatality rates (CFR), t tests and Pearson's χ^2 tests were generated with StataCorp 2013 (Stata Statistical Software, Release 13; College Station, TX).

$$\text{CFR}\ (\%) = \frac{number\ of\ reported\ COVID\text{–}19\ deaths}{number\ of\ reported\ COVID\text{–}19\ cases}\ x\ 100$$

Attribute selection, random undersampling, hyperparameter optimizations, model generation, cross-validation and calculations of model performances were done in RapidMiner Studio 9.10.008 (rev: 68db53, platform: WIN64) (see **Supplementary Information B**). The attribute *RemovalType* was used as the outcome in all data sets. Attribute selection was done individually for all data sets using feature weights operators *weightbyGiniIndex* and *weightbyInformationGainRatio* to determine appropriate attributes to be included in the model generation. Grid optimizations of the hyperparameters for the decision tree operator *decisionTree* for all data sets were done by running fivefold cross-validation using the subprocess *optimizeParameters(Grid)* and operator *crossValidation*. Model generation was done for all four data sets using fivefold cross-validation using the optimized hyperparameters and *RemovalType*=DIED as the positive class set. Random undersampling (RUS) was done only on the training data sets for each fold, and was done using the *sample* operator to a) select all cases with *RemovalType*=DIED, and b) randomly select cases with *RemovalType*=RECOVERED using simple random sampling to achieve a 1:1 RECOVERED:DIED ratio. This training dataset was used to generate the decision tree models per fold. All cases in the testing data sets were used to validate each model.

The decision tree models generated per data set were extracted. Performance metrics such as area under the curve (AUC), accuracy, F-score, sensitivity and specificity were extracted from the cross-validation. Similar cross-validation operators were used to generate naïve Bayes and random forest

models and performance metrics of all data sets. Receiver operating characteristic (ROC) curves for the three models were also generated using RapidMiner Studio 9.10.008 (rev: 68db53, platform: WIN64).

RESULTS

Description of cases

The final data set consisted of 3 896 206 cases (99.02% of all total reported cases from the raw data set) and 10 attributes including one label attribute (*RemovalType*). The A0, AB, D and O data sets comprised 14.68%, 21.45%, 36.31% and 27.56% of the reported cases in the final data set, respectively. The daily reported cases as well as the segmentation according to variants are visualized in **Fig. 1**. Of all reported cases, 98.39% recovered while 1.61% died. Among all reported deaths, the D data set contributed the most cases (43.52%) while the O data set contributed the least (10.36%). Among the four data sets, the highest CFR occurred during the first wave (2.49%) and the lowest during the Omicron wave (0.61%).the Alpha/Beta waves, reported cases were predominantly males, but the CFRs among males were higher than females across all four data sets. Cases with age over 85 years had the highest CFR among different age groups, while cases in the 5-17, 18-29 and 30-39 age groups had the lowest CFRs. Age-stratified CFRs in the Alpha/Beta, Delta and Omicron waves were lower compared to the ancestral wave across age groups (**Table 1**).

Based on disaggregation by region, the National Capital Region (NCR), Cordillera Autonomous Region (CAR), Region II and Region IV-A reported the highest case rates overall (9 304, 7 007, 4 603 and 4 323 cases per 100 000, respectively) and among most of the four data sets. The highest CFR was recorded in Region VII during the Delta wave (4.40%), while the lowest CFR was recorded in repatriated overseas Filipinos (ROF) during the Omicron wave (0.02%) (**Table 1**).

Outcomes from machine learning models

Out of the nine non-outcome attributes retained for model generation, only *Age* and *Admitted* were included in the models for data sets A0, AB and D. For the data set O, *Age*, *Admitted* and *RegionRes* were included in the model. The models were trained and cross-validated with optimized hyperparameters detailed in **Supplementary Information B.3-5**.

In terms of performance, accuracy, F-score and specificity were highest using naïve Bayes in all four data sets. AUC was highest in the naïve Bayes models for the A0, AB and D data sets, while sensitivity was highest in the decision tree models for the A0, AB and O data sets (**Table 2**). The ROC curves for the naïve Bayes and random forest models were better compared to the ROC curve of the decision tree model (**Fig. 2**).

The decision tree models for the A0 and AB data sets were similar: they were composed of three levels, with each node splitting into two branches (**Fig. 3A** and **Fig. 3B**, respectively). The D data set had four levels and had either two or three splits (**Fig. 3C**). The root node for the A0, AB and D datasets was *Age*, with the lowest split criterion in the D data set (41.5 years) and the highest in the A0 data set (47.5 years). Another split according to *Age* was also observed in all three data sets at *Age* = 0.5 years. The attribute *Admitted* also split the D data set for cases with *Age* <= 41.5 years (**Fig. 3C**). Majority of cases above the root node cutoffs died in all three data sets (A0 = 76.60%, AB = 72.93%, D = 70.21%), while majority of cases within or below the root node cutoff and above *Age* = 0.5 years recovered (A0 = 82.88%, AB = 86.19%, D = 86.39%). In the D data set, 64.04% of cases who had a history of hospital admission died. In the A0, AB and D data sets, majority of cases below *Age* = 0.5 years died (A0 = 65.88%, AB = 59.70%, D = 55.61%).

The O data set had eight levels, but the number of node splits ranged between two and 16 (**Fig. 3D**). The root node was *Age* with a split criterion of 52.5 years. Cases with *Age* <= 52.5 years were further split according to *Age* <= 41.5 years, with 77.02% of those less than 41.5 years recovering. Cases with

Age between 41.5 and 52.5 were split into their region of residence, with the majority outcome = DIED for those residing in Regions I, II, III, IV-B, VI, VII, XI, XII, XIII as well as in CAR. Cases with *Age* > 52.5 years were split into region of residence, with majority of cases from any region dying except for repatriate overseas Filipinos. **Supplementary Information B.5** provides the full information of the decision tree models, including the actual number of cases and outcomes per leaf.

DISCUSSION

We generated four different decision tree models corresponding to the different predominant COVID-19 strain and variants in the Philippines, with age being the root node for all models. The A0 and AB data sets generated simple and similar decision trees with only age as the significant attribute, while the D data set model incorporated admission history as an additional attribute. The O data set generated a more complicated decision tree which incorporated age, admission history and region of residence of the cases into the model. Machine learning models such as decision trees have been used in analyzing trends in COVID-19 data, including in epidemiological modeling[21] and prediction of disease prognosis.[17,20]

Reported COVID-19 cases in the Philippines reached almost 4 million cases as of September 24, 2022, with most cases occurring during the Delta and Omicron waves despite the relatively shorter duration of these waves compared with the first infection wave from the ancestral strain. SARS-CoV-2 variants have shown increasing transmissibility compared to previous ones, with the Delta and Omicron variants reaching R_0 of 7 and 10, respectively, compared to 2.5 of the ancestral strain.[8,22] Other studies reported that the Omicron variant was up to 3.7 times more transmissible compared to the Delta variant and is primarily due to its ability for immune evasion and reinfection regardless of vaccination status and previous infection[23,24] mainly due to an enhanced viral replication efficiency in the bronchus.[25]

Previous studies have found that the severity among the ancestral, Alpha and Delta variants are comparable,[11,18] but the severity of the Omicron variant has consistently been lower compared to the other variants.[11,24,26,27] This may be due to lower replication competence of the Omicron variant in the lung parenchyma.[25] These findings were consistent with our study: our calculated CFR during the Omicron wave was 65%, 68% and 75% lower than those of the Alpha/Beta, Delta and ancestral waves, respectively. Earlier studies on sex differentials in COVID-19 mortality (i.e., males tend to have higher CFRs)[20,28] also confirm our results regardless of COVID-19 variant.

In our study, age is the main predictor of our defined outcome for reported COVID-19 cases. Older age groups tend to have higher case fatality rates regardless of predominant COVID-19 variant. This general trend has been documented in previous studies.[1,6,11,29,30] However, we noticed a pattern similar to a previous Philippine study[20] on cases of the ancestral variant: the CFRs of the lowest age group (i.e., 0-4 years) tend to be up to 6 times the CFR of the baseline (i.e., 18-29 years), with the lowest CFRs seen in the 5-17 age group. The US Centers for Disease Control and Prevention[29] shows a generally increasing trend in CFR but a study by Khera et al. (2021) supports our findings and attributed this “U-shaped” phenomenon to several factors such as children having differential expression of ACE-2 receptors, more robust innate immune system (except for newborns), and lesser exposure due to public health measures.[31]

Our decision tree models showed several results. First, among all the attributes included in our models and consistent with our descriptive analysis, age is the most important predictor of mortality. Previous machine learning models on COVID-19 mortality[20,32] confirm this finding, suggesting that in the absence of clinical data in surveillance data sets, age remains an important factor. Second, the similarities between the A0 and AB models suggest that earlier in the pandemic, the impact of the two waves in the general population may have been similar. During these times, large portions of the population in the country were still under COVID-19 lockdowns and vaccinations had barely started.[33,34] These events may have limited the population’s exposure to the virus and to COVID-19 vaccines which may suggest that during the early months of the pandemic, internal biological factors such as age-related immunosenescence and presence of comorbidities are bigger factors in prognosis compared to natural or acquired immunity.[1] Third, the D model incorporated history of admission as a splitting criterion,

similar to a previous study.[20] The previous national guidelines for hospitalization of COVID-19 patients prioritizes admission of only severe and critical COVID-19 cases,[20,35] and this may have been exacerbated by the sudden influx of COVID-19 cases during the Delta wave as reported in this study.

Fourth, the incorporation of the attribute *RegionRes* (region of residence) in the O data set model is quite novel. A previous study[20] of the early COVID-19 ancestral wave in the Philippines did not include geopolitical classifications in the model and was consistent with our A0, AB and D models. Our current O model suggests that there may be different impacts of the Omicron variant among different regions in the Philippines. Literature regarding regional differences in COVID-19 CFRs are limited, but previous studies recognized the association of transmission or mortality rates with differences in health care system factors such as number of available hospital beds,[10,36,37] length and severity of lockdowns, population or industrial composition,[37,38] and previous infection or vaccination rates.[5,7,12,27] Repatriated overseas Filipinos, on the other hand, are only allowed to return to the country if they are well enough to travel,[39] hence the lower CFR among this cohort regardless of infection wave.

Our study has several limitations. Since the data set is publicly available surveillance data, it did not include clinical factors that are associated with COVID-19 mortality such as presence of comorbidities, vaccination status and sociodemographic information. Our categorization of cases according to infection waves was also based on the predominant variant during the date of confirmation of infection and not based on genetic sequencing. Additionally, these dates may have also been delayed. These factors could have led to improper classification of reported cases, particularly those whose reported dates were near the boundaries of our infection wave timelines.

Surveillance data sets during the pandemic are often imbalanced in that the number of recoveries vastly outnumber reported deaths. We used undersampling techniques to control for this imbalance. The models we generated generally had high AUC and sensitivity, with the naïve Bayes and the decision tree models mostly having the highest AUC and sensitivity across the different data sets, respectively. Higher sensitivity is often preferred in inherently imbalanced data sets.[40] We utilized similar techniques from a previous study[20] to reduce overfitting: removing irrelevant or highly correlated attributes, enabling pre-pruning and pruning during model training, and optimizing the hyperparameters for the highest sensitivity.

In conclusion, our study highlights the observable changes in COVID-19 transmissibility and case fatality rates depending on the infection timeline and predominant SARS-CoV-2 variant, with the mortality pattern of the Omicron variant being significantly different from the preceding variants. Our findings also reinforce the strong influence of increasing age in predicting COVID-19 outcomes regardless of SARS-CoV-2 variant. The models that we generated highlight the need for up-to-date and stratified policies especially during viral epidemics and pandemics. We recommend future research to apply similar analysis of publicly available surveillance data to monitor emerging or ongoing outbreaks.

Acknowledgements

The authors would like to thank the San Beda Research and Development Center for the overall support to the study.

Conflicts of interest

The authors have no conflicts of interest to declare.

Ethics statement

The study was reviewed and approved by the San Beda University Research Ethics Board on October 21, 2022 under the study protocol code SBU-RED 2022-020. The study adhered to the TRIPOD checklist for prediction model development.

Funding

The study was funded in part by an operational grant from the San Beda University Office of Research and Innovation under the study protocol code SBU-REB 2022-020.

REFERENCES

1. Malundo AFG, Abad CLR, Salamat MSS, Sandejas JCM, Poblete JB, Planta JEG, et al. Predictors of mortality among inpatients with COVID-19 infection in a tertiary referral center in the Philippines. IJID Regions. 2022 Sep 1;4:134–42.

2. DOH. COVID-19 Tracker | Department of Health website [Internet]. 2022 [cited 2022 Oct 8]. Available from: https://doh.gov.ph/covid19tracker

3. WHO. WHO Coronavirus Disease (COVID-19) Dashboard [Internet]. WHO Coronavirus Disease (COVID-19) Dashboard. 2023 [cited 2022 Dec 1]. Available from: https://covid19.who.int

4. Johnson AG, Amin A, Ali A, et al. COVID-19 Incidence and Death Rates Among Unvaccinated and Fully Vaccinated Adults with and Without Booster Doses During Periods of Delta and Omicron Variant Emergence — 25 U.S. Jurisdictions, April 4–December 25, 2021. MMWR Morb Mortal Wkly Rep [Internet]. 2022 [cited 2022 Oct 8];71. Available from: https://www.cdc.gov/mmwr/volumes/71/wr/mm7104e2.htm

5. Stein C, Nassereldine H, Sorensen RJD, Amlag JO, Bisignano C, Byrne S, et al. Past SARS-CoV-2 infection protection against re-infection: a systematic review and meta-analysis. The Lancet [Internet]. 2023 Feb 16 [cited 2023 Feb 17];0(0). Available from: https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(22)02465-5/fulltext?dgcid=raven_jbs_etoc_feature_lancet

6. SeyedAlinaghi S, Mirzapour P, Dadras O, Pashaei Z, Karimi A, MohsseniPour M, et al. Characterization of SARS-CoV-2 different variants and related morbidity and mortality: a systematic review. European Journal of Medical Research. 2021 Jun 8;26(1):51.

7. Bhattacharyya RP, Hanage WP. Challenges in Inferring Intrinsic Severity of the SARS-CoV-2 Omicron Variant. New England Journal of Medicine. 2022 Feb 17;386(7):e14.

8. Lorente-González M, Suarez-Ortiz M, Landete P. Evolution and Clinical Trend of SARS-CoV-2 Variants. Open Respiratory Archives. 2022 Apr 1;4(2):100169.

9. WHO. Tracking SARS-CoV-2 variants [Internet]. World Health Organization. 2022 [cited 2022 Oct 8]. Available from: https://www.who.int/activities/tracking-SARS-CoV-2-variants

10. Carbonell R, Urgelés S, Rodríguez A, Bodí M, Martín-Loeches I, Solé-Violán J, et al. Mortality comparison between the first and second/third waves among 3,795 critical COVID-19 patients with pneumonia admitted to the ICU: A multicentre retrospective cohort study. Lancet Reg Health Eur. 2021 Dec;11:100243.

11. Esper FP, Adhikari TM, Tu ZJ, Cheng YW, El-Haddad K, Farkas DH, et al. Alpha to Omicron: Disease Severity and Clinical Outcomes of Major SARS-CoV-2 Variants. The Journal of Infectious Diseases. 2023 Feb 1;227(3):344–52.

12. Kläser K, Molteni E, Graham M, Canas LS, Österdahl MF, Antonelli M, et al. COVID-19 due to the B.1.617.2 (Delta) variant compared to B.1.1.7 (Alpha) variant of SARS-CoV-2: a prospective observational cohort study. Sci Rep. 2022 Jun 28;12(1):10904.

13. Christensen PA, Olsen RJ, Long SW, Snehal R, Davis JJ, Saavedra MO, et al. Signals of Significantly Increased Vaccine Breakthrough, Decreased Hospitalization Rates, and Less Severe Disease in Patients with Coronavirus Disease 2019 Caused by the Omicron Variant of Severe Acute Respiratory Syndrome Coronavirus 2 in Houston, Texas. The American Journal of Pathology. 2022 Apr 1;192(4):642–52.

14. Re3data.Org: GISAID. Global COVID-19 submission tracker [Internet]. GISAID. re3data.org - Registry of Research Data Repositories; 2022 [cited 2022 Oct 8]. Available from: https://gisaid.org/submission-tracker-global/

15. Re3data.Org: GISAID. hCoV-19 Variants Dashboard [Internet]. GISAID. re3data.org - Registry of Research Data Repositories; 2022 [cited 2022 Oct 8]. Available from: https://gisaid.org/hcov-19-variants-dashboard/

16. Painuli D, Mishra D, Bhardwaj S, Aggarwal M. Forecast and prediction of COVID-19 using machine learning. Data Science for COVID-19. 2021;381–97.

17. Mahdavi M, Choubdar H, Zabeh E, Rieder M, Safavi-Naeini S, Jobbagy Z, et al. A machine learning based exploration of COVID-19 mortality risk. PLOS ONE. 2021 Jul 2;16(7):e0252384.

18. Hu Z, Huang X, Zhang J, Fu S, Ding D, Tao Z. Differences in Clinical Characteristics Between Delta Variant and Wild-Type SARS-CoV-2 Infected Patients. Frontiers in Medicine [Internet]. 2022 [cited 2023 Feb 5];8. Available from: https://www.frontiersin.org/articles/10.3389/fmed.2021.792135

19. Noy O, Coster D, Metzger M, Atar I, Shenhar-Tsarfaty S, Berliner S, et al. A machine learning model for predicting deterioration of COVID-19 inpatients. Sci Rep. 2022 Feb 16;12(1):2630.

20. Migriño JR, Batangan ARU. Using machine learning to create a decision tree model to predict outcomes of COVID-19 cases in the Philippines: Decision tree for COVID-19 cases. Western Pacific Surveillance and Response. 2021 Sep 14;12(3):9–9.

21. Venkatasubramaniam A, Wolfson J, Mitchell N, Barnes T, JaKa M, French S. Decision trees in epidemiological research. Emerging Themes in Epidemiology. 2017 Sep 20;14(1):11.

22. Rashedi R, Samieefar N, Akhlaghdoust M, Mashhadi M, Darzi P, Rezaei N. Delta Variant: The New Challenge of COVID-19 Pandemic, an Overview of Epidemiological, Clinical, and Immune Characteristics. Acta Biomed. 2022;93(1):e2022179.

23. Mohsin M, Mahmud S. Omicron SARS-CoV-2 variant of concern: A review on its transmissibility, immune evasion, reinfection, and severity. Medicine. 2022 May 13;101(19):e29165.

24. Wolter N, Jassat W, Walaza S, Welch R, Moultrie H, Groome M, et al. Early assessment of the clinical severity of the SARS-CoV-2 omicron variant in South Africa: a data linkage study. The Lancet. 2022 Jan 29;399(10323):437–46.

25. Hui KPY, Ho JCW, Cheung M chun, Ng K chun, Ching RHH, Lai K ling, et al. SARS-CoV-2 Omicron variant replication in human bronchus and lung ex vivo. Nature. 2022 Mar;603(7902):715–20.

26. Lewnard JA, Hong VX, Patel MM, Kahn R, Lipsitch M, Tartof SY. Clinical outcomes associated with SARS-CoV-2 Omicron (B.1.1.529) variant and BA.1/BA.1.1 or BA.2 subvariant infection in Southern California. Nat Med. 2022 Sep;28(9):1933–43.

27. Wang C, Liu B, Zhang S, Huang N, Zhao T, Lu Q, et al. Differences in incidence and fatality of COVID-19 by SARS-CoV-2 Omicron variant versus Delta variant in relation to vaccine coverage: A world-wide review. J Med Virol. 2022 Sep 20;10.1002/jmv.28118.

28. Bhopal SS, Bhopal R. Sex differential in COVID-19 mortality varies markedly by age. The Lancet. 2020 Aug 22;396(10250):532–3.

29. CDC. Risk for COVID-19 Infection, Hospitalization, and Death By Age Group [Internet]. Centers for Disease Control and Prevention. 2022 [cited 2023 Feb 6]. Available from: https://www.cdc.gov/coronavirus/2019-ncov/covid-data/investigations-discovery/hospitalization-death-by-age.html

30. Endeshaw Y, Campbell K. Advanced age, comorbidity and the risk of mortality in COVID-19 infection. J Natl Med Assoc. 2022 Oct;114(5):512–7.

31. Khera N, Santesmasses D, Kerepesi C, Gladyshev VN. COVID-19 mortality rate in children is U-shaped. Aging (Albany NY). 2021 Aug 18;13(16):19954–62.

32. Yadaw AS, Li Y chak, Bose S, Iyengar R, Bunyavanich S, Pandey G. Clinical features of COVID-19 mortality: development and validation of a clinical prediction model. The Lancet Digital Health. 2020 Oct;2(10):e516–25.

33. Argosino F. COVID-19 response: A timeline of community quarantine, lockdowns, alert levels [Internet]. Manila Bulletin. 2021 [cited 2022 Aug 6]. Available from: https://mb.com.ph/2021/11/09/covid-19-response-a-timeline-of-community-quarantine-lockdowns-alert-levels/

34. DOH. Updates on COVID-19 Vaccines | COVID-19 Vaccination Dashboard [Internet]. 2023 [cited 2023 Feb 10]. Available from: https://doh.gov.ph/vaccines

35. DOH. Interim Guidelines on the COVID-19 Disease Severity Classification and Management [Internet]. Manila, Philippines: Department of Health; 2020 Jul. Report No.: Department Memorandum 2020-0381. Available from: https://doh.gov.ph/node/24520

36. Pan J, St. Pierre JM, Pickering TA, Demirjian NL, Fields BKK, Desai B, et al. Coronavirus Disease 2019 (COVID-19): A Modeling Study of Factors Driving Variation in Case Fatality Rate by Country. Int J Environ Res Public Health. 2020 Nov;17(21):8189.

37. Talabis DAS, Babierra AL, Buhat CAH, Lutero DS, Quindala KM, Rabajante JF. Local government responses for COVID-19 management in the Philippines. BMC Public Health. 2021 Dec;21(1):1–15.

38. Jiang Y, Laranjo JR, Thomas M. COVID-19 Lockdown Policy and Heterogeneous Responses of Urban Mobility: Evidence from the Philippines [Internet]. Asian Development Bank; 2022 May [cited 2023 Mar 3]. (Economics Working Papers). Available from: https://www.adb.org/publications/covid-19-lockdown-policy-urban-mobility-philippines

39. IATF. IATF RESOLUTION NO. 168 [Internet]. 2022 May [cited 2023 Mar 7]. Report No.: 168. Available from: https://iatf.doh.gov.ph/2022/05/26/iatf-resolution-no-168/

40. Serrano LG. Grokking machine learning. Shelter Island: Manning Publications; 2021. 489 p.

Figure 1. Reported COVID-19 cases in the Philippines by predominant variant

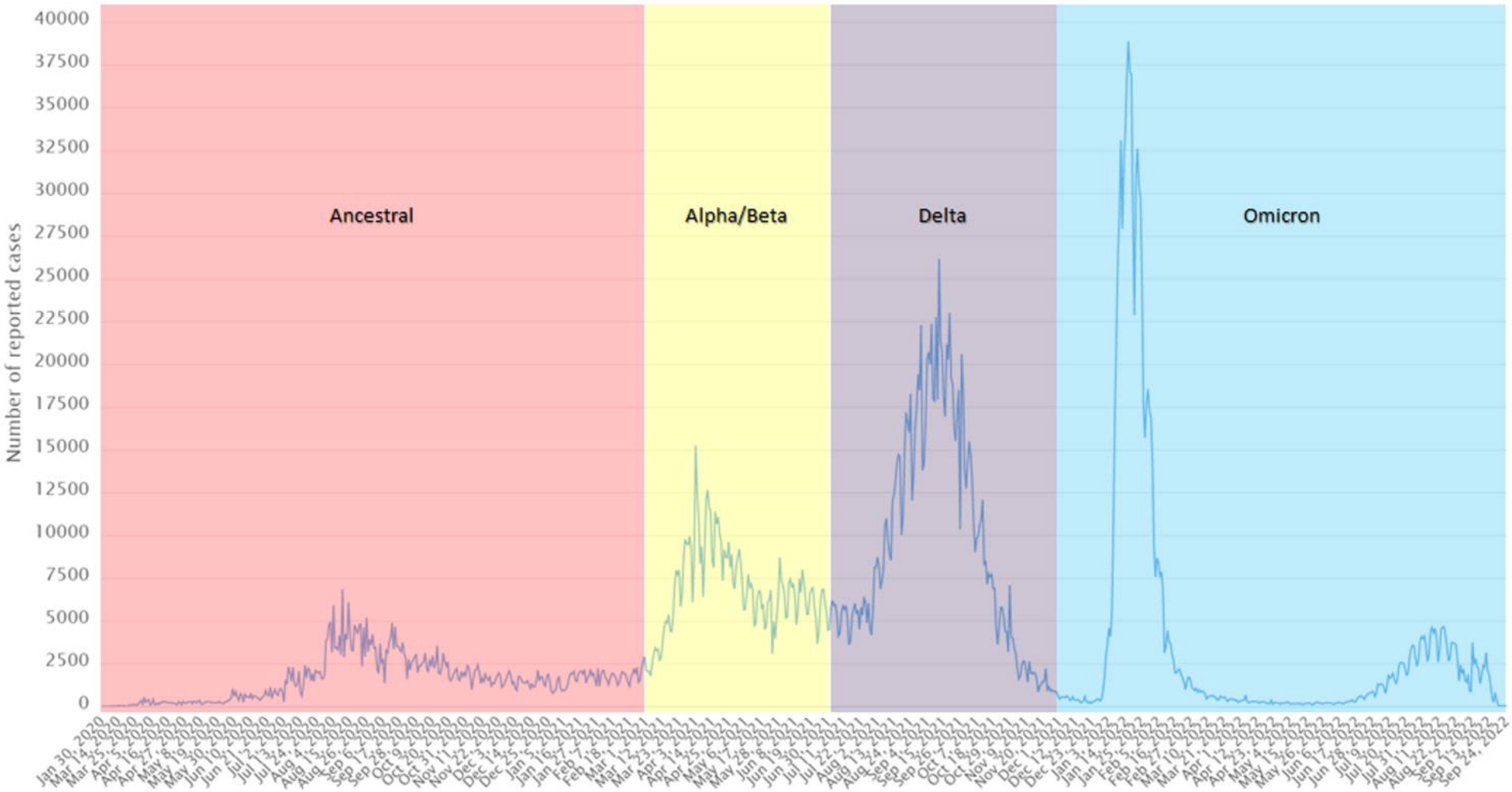
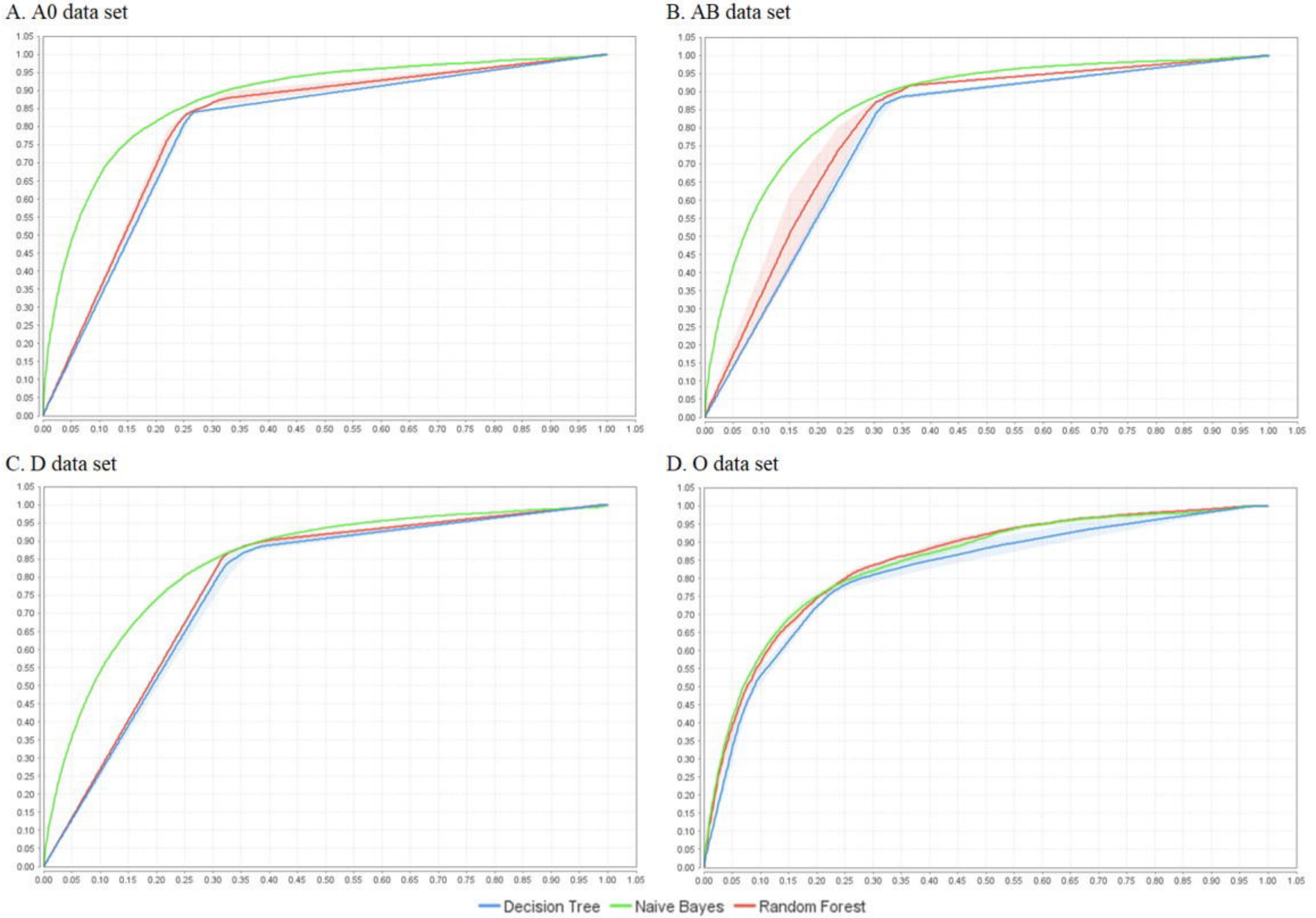


Table 1. Demographic characteristics of reported cases (recovered or died) from the Philippines COVID Data Drop from September 24, 2022

	Overall			Ancestral variant			Alpha/Beta variant			Delta variant			Omicron variant		
	Recovered	**Died**	**CFR (%)**	**Recovered**	**Died**	**CFR (%)**	**Recovered**	**Died**	**CFR (%)**	**Recovered**	**Died**	**CFR (%)**	**Recovered**	**Died**	**CFR (%)**
Overall	3 833 494	62 712	1.61%	557 657	14 248	2.49%	821 104	14 675	1.76%	1 387 359	27 292	1.93%	1 067 374	6 497	0.61%
Sex	*n = 3 896 206*			*n = 571 905*			*n = 835 779*			*n = 1 414 651*			*n = 1 073 871*		
Male	1 865 640	34 634	1.82%	298 688	8 571	2.79%	415 121	8 080	1.91%	664 988	14 354	2.11%	486 843	3 629	0.74%
Female	1 967 854	28 078	1.41%	258 969	5 677	2.15%	405 983	6 595	1.60%	722 371	12 938	1.76%	580 531	2 868	0.49%
Age	*n = 3 896 206*			*n = 571 905*			*n = 835 779*			*n = 1 414 651*			*n = 1 073 871*		
Mean age, *years (S.D.)*	37.39 (±17.68)	62.04 (±17.41)		37.61 (±16.49)	62.15 (±16.60)		38.20 (±17.58)	63.45 (±15.81)		37.48 (±18.68)	62.06 (±17.39)		36.54 (±16.96)	58.52 (±21.65)	
Age Group	*n = 3 896 206*			*n = 571 905*			*n = 835 779*			*n = 1 414 651*			*n = 1 073 871*		
0-4	92 836	715	0.76%	9 477	144	1.50%	15 843	87	0.55%	37 737	271	0.71%	29 779	213	0.71%
5-17	289 366	510	0.18%	32 327	93	0.29%	59 509	77	0.13%	135 749	198	0.15%	61 781	142	0.23%
18-29	1 042 946	1 897	0.18%	157 193	413	0.26%	217 215	309	0.14%	354 779	806	0.23%	313 759	369	0.12%
30-39	891 547	3 327	0.37%	135 646	681	0.50%	184 552	638	0.34%	294 524	1 533	0.52%	276 825	475	0.17%
40-49	587 246	6 063	1.02%	91 350	1 305	1.41%	129 162	1 338	1.03%	203 023	2 803	1.36%	163 711	617	0.38%
50-64	610 242	19 258	3.06%	92 717	4 568	4.70%	143 444	4 574	3.09%	229 380	8 351	3.51%	144 701	1 765	1.21%
65-74	205 764	16 127	7.27%	27 149	3 945	12.69%	48 514	4 180	7.93%	83 848	6 643	7.34%	46 253	1 359	2.85%
75-84	85 606	10 380	10.81%	9 297	2 283	19.72%	17 879	2 458	12.09%	36 445	4 636	11.29%	21 985	1 003	4.36%
85+	27 941	4 435	13.70%	2 501	816	24.60%	4 986	1 014	16.90%	11 874	2 051	14.73%	8 580	554	6.07%
Region	*n = 3 892 573*			*n = 568 966*			*n = 835 727*			*n = 1 414 514*			*n = 1 073 366*		
NCR	1 233 773	13 250	1.05%	227 269	5 426	2.33%	287 973	3 620	1.24%	328 595	3 022	0.91%	389 936	1 182	0.30%
Region I	138 205	2 706	1.88%	8 432	315	3.60%	19 764	510	2.52%	74 242	1 491	1.97%	35 767	390	1.08%
Region II	164 646	4 732	2.72%	9 030	190	2.06%	42 613	1 210	2.76%	81 455	2 860	3.39%	31 548	472	1.47%
Region III	376 322	7 742	1.98%	36 626	1 267	3.34%	86 141	2 276	2.57%	151 199	3 225	2.09%	102 356	974	0.94%
Region IV-A	691 369	6 415	0.91%	95 727	1 731	1.78%	143 466	1 715	1.18%	250 897	2 409	0.95%	201 279	560	0.28%
Region IV-B	45 204	1 254	2.63%	2 943	66	2.19%	10 940	387	3.42%	22 247	680	2.97%	9 074	121	1.32%
Region V	68 654	1 171	1.65%	5 731	207	3.49%	14 396	246	1.68%	30 438	580	1.87%	18 089	138	0.76%
Region VI	201 994	5 503	2.58%	25 491	801	3.05%	42 020	1 128	2.61%	79 022	2 973	3.63%	55 461	601	1.07%
Region VII	195 487	6 394	3.07%	38 472	1 714	4.27%	35 293	630	1.75%	74 122	3 408	4.40%	47 600	642	1.33%
Region VIII	65 226	861	1.29%	15 621	239	1.51%	14 626	244	1.64%	22 080	294	1.31%	12 899	84	0.65%
Region IX	67 192	1 456	2.08%	7 102	270	3.66%	17 082	475	2.71%	27 614	610	2.16%	15 394	101	0.65%
Region X	108 663	1 132	1.02%	11 530	292	2.47%	19 864	256	1.27%	52 928	504	0.94%	24 341	80	0.33%
Region XI	142 457	3 920	2.61%	19 788	836	4.05%	21 644	572	2.57%	61 360	2 065	3.26%	39 665	447	1.11%
Region XII	77 689	1 317	1.64%	5 835	188	3.12%	15 904	341	2.10%	36 917	651	1.73%	19 033	137	0.71%
Region XIII	62 385	1 711	2.60%	7 398	289	3.76%	14 238	344	2.36%	27 424	890	3.14%	13 325	188	1.39%
BARMM	26 023	594	2.18%	4 319	133	2.99%	5 618	156	2.70%	9 122	256	2.73%	6 964	49	0.70%
CAR	123 060	2 450	1.91%	14 262	228	1.57%	24 535	551	2.20%	51 666	1 342	2.53%	32 597	329	1.00%
ROF	41 520	96	0.23%	19 150	48	0.25%	4 935	14	0.28%	5 894	32	0.54%	11 541	2	0.02%

BARMM: Bangsamoro Autonomous Region in Muslim Mindanao; CAR: Cordillera Administrative Region; NCR: National Capital Region; ROF: repatriated overseas Filipinos

Figure 2. Receiver operating characteristic (ROC) curves for the machine learning models by data set: decision tree, naïve Bayes, random forest[a]



[a] The ROC curve plots a model's sensitivity, or true positive rate, versus its false positive rate (one minus the specificity or true negative rate) as its discrimination threshold is varied. Generally, the closer the ROC curve is to the top left corner of the graph, the better the model.

Table 2. Performance metrics for the three machine learning models: decision tree, naïve Bayes and random forest using the four modelling data sets and optimized hyperparameters

A0 dataset										
Model	**AUC**		**Accuracy**		**F-score**		**Sensitivity**		**Specificity**	
Decision Tree	0.789	± 0.004	74.06%	± 0.79%	13.88%	± 0.35%	83.86%[a]	± 0.35%	73.81%	± 0.81%
Naïve Bayes	0.877[a]	± 0.004	80.25%[a]	± 0.60%	17.00%[a]	± 0.36%	81.16%	± 0.55%	80.23%[a]	± 0.62%
Random Forest	0.824	± 0.018	74.66%	± 0.62%	14.10%	± 0.28%	83.43%	± 0.54%	74.43%	± 0.65%

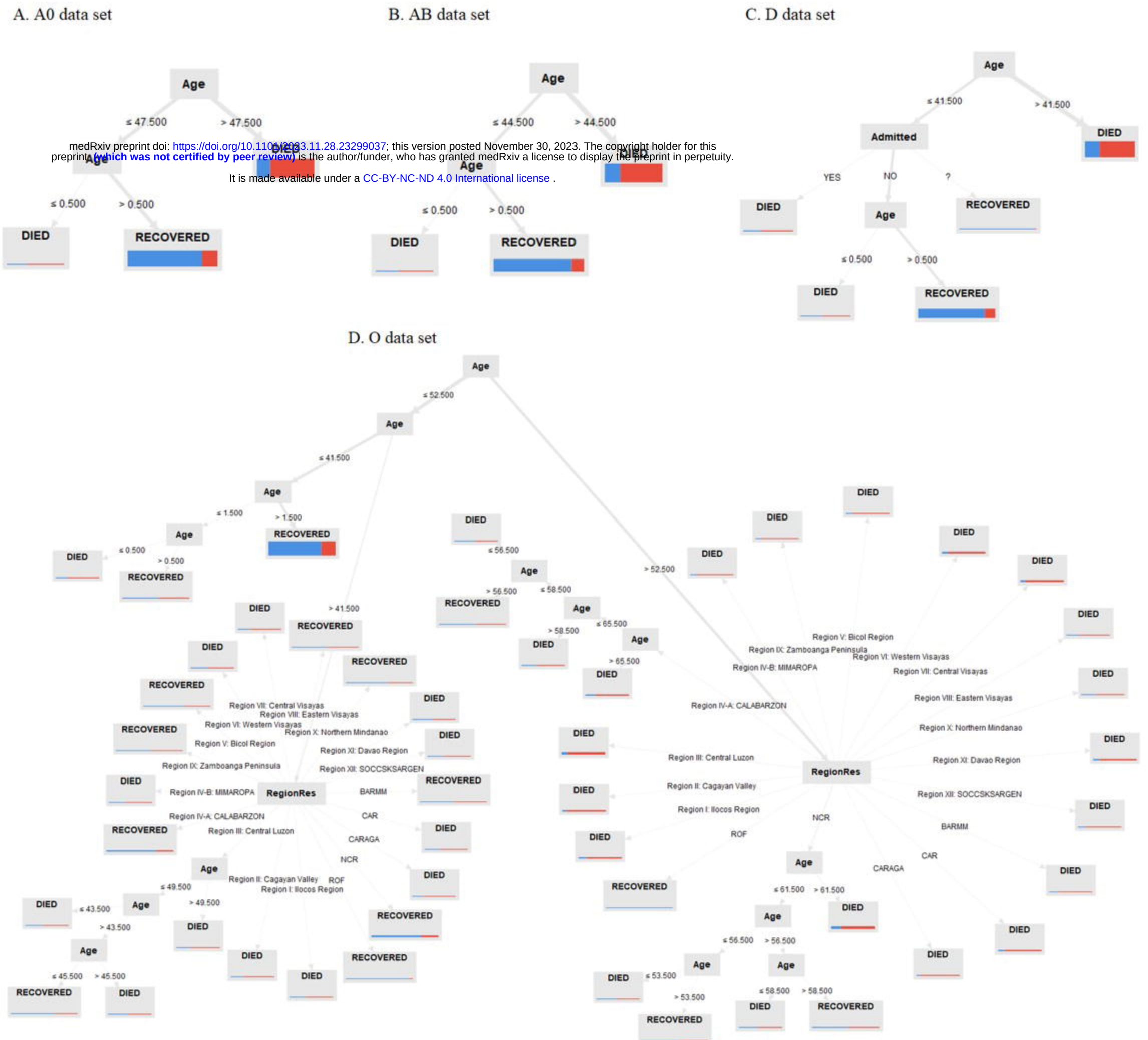
AB dataset										
Model	**AUC**		**Accuracy**		**F-score**		**Sensitivity**		**Specificity**	
Decision Tree	0.781	± 0.004	68.36%	± 1.80%	8.91%	± 0.35%	88.03%[a]	± 1.21%	68.00%	± 1.85%
Naïve Bayes	0.869[a]	± 0.004	77.07%[a]	± 0.08%	11.22%[a]	± 0.11%	82.51%	± 0.62%	76.97%[a]	± 0.07%
Random Forest	0.798	± 0.003	68.76%	± 1.11%	8.99%	± 0.22%	87.79%	± 0.90%	68.42%	± 1.14%

D dataset										
Model	**AUC**		**Accuracy**		**F-score**		**Sensitivity**		**Specificity**	
Decision Tree	0.769	± 0.006	66.15%	± 2.73%	9.12%	± 0.48%	87.69%	± 1.77%	65.73%	± 2.82%
Naïve Bayes	0.844[a]	± 0.003	74.74%[a]	± 0.05%	10.98%[a]	± 0.08%	80.73%	± 0.58%	74.62%[a]	± 0.05%
Random Forest	0.779	± 0.005	65.32%	± 2.62%	8.96%	± 0.46%	88.18%[a]	± 1.73%	64.87%	± 2.70%

O dataset										
Model	**AUC**		**Accuracy**		**F-score**		**Sensitivity**		**Specificity**	
Decision Tree	0.814	± 0.014	77.27%	± 2.34%	3.93%	± 0.26%	76.42%[a]	± 2.82%	77.28%	± 2.37%
Naïve Bayes	0.843	± 0.006	80.30%[a]	± 0.24%	4.38%[a]	± 0.08%	74.53%	± 1.55%	80.33%[a]	± 0.24%
Random Forest	0.844[a]	± 0.006	78.32%	± 1.02%	4.09%	± 0.13%	76.25%	± 1.61%	78.33%	± 1.04%

[a] Highest values for each metric across all models

Figure 3. Decision tree for predicted outcomes of reported cases (recovered or died) by data set from the Philippines COVID Data Drop from September 24, 2022[a]



[a] Relevant attributes identified by the model are shown inside the branches. The predominant outcome per leaf node is identified (either RECOVERED or DIED), with the coloured bars underneath illustrating horizontal stacked bars of the predominant outcome per leaf (RECOVERED=blue, DIED=red). The width of the bars represents the relative number of cases in each leaf as compared with the total cases in the modeling dataset, while the thickness of each arrow illustrates the relative number of cases on each branch as compared with the total cases in the modeling dataset.