Diversity: Disease Preventive of Research Integrity
Volume 5, Issue 1, 2024
Page 10-21
doi: 10.24252/diversity.v5i1.49508

p-ISSN : 2774-9649
e-ISSN : 2775-0035

COVID-19 Infection Wave Mortality from Surveillance Data in The Philippines Using Machine Learning

Kematian Gelombang Infeksi COVID-19 dari Data Surveilans di Filipina Menggunakan Pembelajaran Mesin

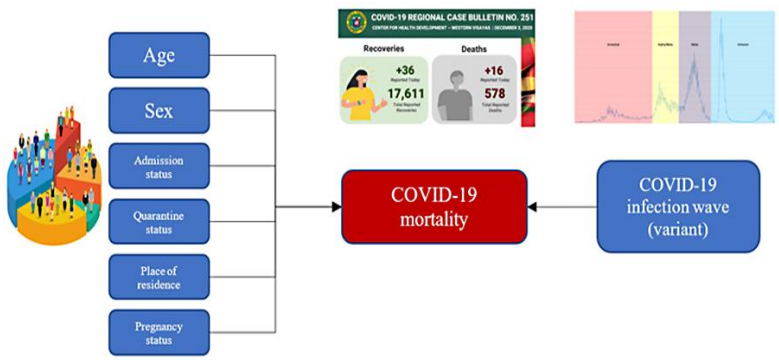
Julius R. Migriño Jr*[1, 4], Ani R. U. Batangan[2], Rizal M. R. Abello[3]

[1, 2, 3] College of Medicine, San Beda University, Manila, Philippines
[4] School of Medicine and Public Health, Ateneo de Manila University, Pasig, Philippines

Abstract
The Philippines 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. A cross-sectional surveillance data review design was employed using the DOH COVID Data Drop dataset as of September 24, 2022. The predominant variant(s) of concern divided the dataset into time intervals representing the infection waves: ancestral (A0), Alpha/Beta (AB), Delta (D), and Omicron (O). Descriptive statistics and machine learning models were generated from each infection. The final data set consisted of 3,896,206 cases wherein 98.39% of cases recovered while 1.61% died. The highest and lowest CFR was observed during the ancestral wave (2.49) and the Omicron wave (0.61%), respectively. In all four data sets, higher age groups had higher CFRs, and F-score and specificity were highest using naïve Bayes. 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. The ancestral, Alpha/Beta and Delta variants seem to have similar transmission and mortality profiles, while the Omicron variant caused lesser deaths despite increased transmissibility.

Abstrak
Filipina memiliki beberapa gelombang infeksi COVID-19 yang disebabkan oleh strain dan varian SARS-CoV-2 yang berbeda. Penelitian ini bertujuan untuk menggambarkan hasil COVID-19 berdasarkan gelombang infeksi menggunakan pembelajaran mesin. Desain tinjauan data surveilans cross-sectional digunakan dengan menggunakan set data DOH COVID Data Drop pada 24 September 2022. Varian utama yang menjadi perhatian membagi dataset ke dalam interval waktu yang mewakili gelombang infeksi: leluhur (A0), Alpha/Beta (AB), Delta (D), dan Omicron (O). Statistik deskriptif dan model pembelajaran mesin dihasilkan dari setiap infeksi. Kumpulan data akhir terdiri dari 3.896.206 kasus di mana 98,39% kasus sembuh dan 1,61% meninggal. CFR tertinggi dan terendah diamati selama gelombang leluhur (2,49) dan gelombang Omicron (0,61%). Pada keempat set data, kelompok usia yang lebih tinggi memiliki CFR yang lebih tinggi, dan skor-F dan spesifisitas tertinggi menggunakan naïve Bayes. Area di bawah kurva (AUC) tertinggi dalam model naïve Bayes untuk model A0, AB dan D, sementara sensitivitas tertinggi dalam model pohon keputusan untuk model A0, AB dan O. Varian leluhur, Alpha/Beta dan Delta tampaknya memiliki profil penularan dan kematian yang serupa, sementara varian Omicron menyebabkan kematian yang lebih rendah meskipun penularannya meningkat.

Graphical Abstract



Keyword
covid-19; decision trees; machine learning; mortality; surveillance

Artikel History
Submited : 11 July 2024
In Reviewed : 07 August 2024
Accepted : 28 August 2024
Published : 30 August 2024

Correspondence
Address : 638 Mendiola Street, San Miguel, Manila 1000, Philippines
Email : jrmjrmd-1@yahoo.com

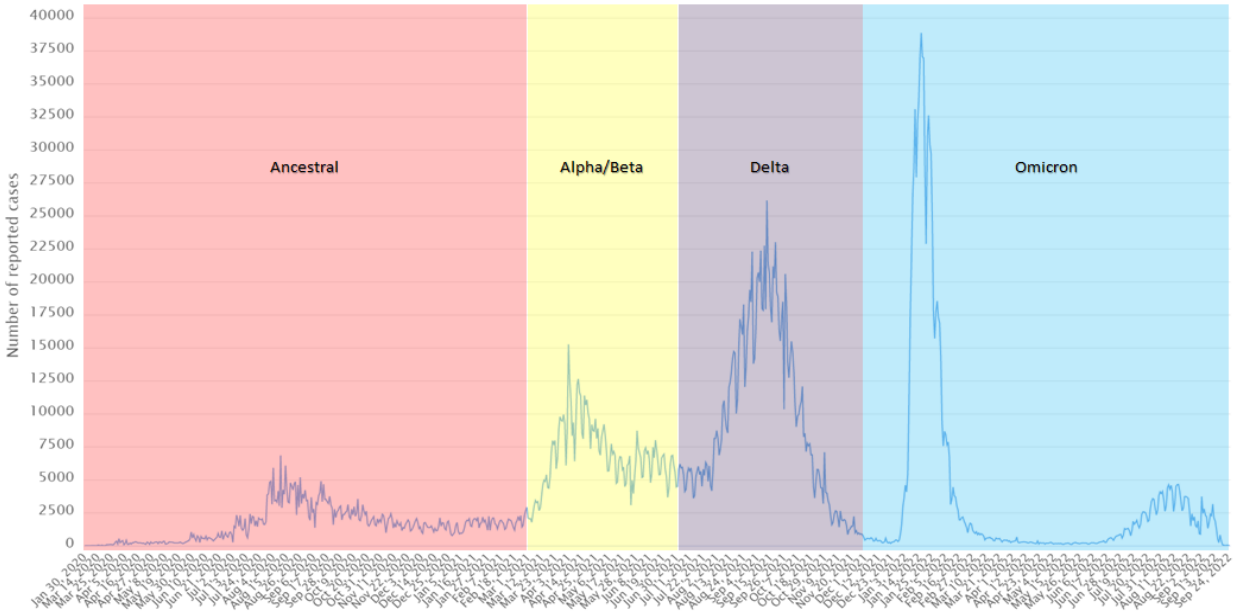
INTRODUCTION

The Philippines has been considered a hotspot for the coronavirus disease 2019 (COVID-19) in the Western Pacific region (Malundo et al., 2022). 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 (DOH, 2022). Meanwhile, the World Health Organization (WHO) has tallied 639,572,819 confirmed cases and 6,615,258 deaths globally (WHO, 2023). 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 life-threatening 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 (Malundo et al., 2022). More recently, studies have surfaced highlighting the differences in mortality rates among cases with different vaccination statuses (Johnson et al., 2022; Stein et al., 2023) and among those who were previously infected (Stein et al., 2023). According to SeyedAlighani et al. (2021), mortality rates are additionally influenced by adequacy of health care delivery, political decisions, and epidemiological characteristics of the affected population. Generally, viruses evolve to become more transmissible, regardless of severity (Bhattacharyya & Hanage, 2022). 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 (Lorente-González et al., 2022). As of December 2022, there had been five recognized circulating SARS-CoV-2 variants of concern (VOCs): Alpha, Beta, Gamma, Delta and Omicron. These VOCs appeared in infection waves among different countries in varying timelines. WHO designated them as VOCs on December 2020 (Alpha and Beta), January 2021 (Gamma), May 2021 (Delta) and November 2021 (Omicron) (WHO, 2022). Recent studies characterized the different VOCs in terms of their transmissibility and severity. For instance, while the Delta variant evolved to become more transmissible, several studies report similar hospitalization and mortality rates among the different infection waves (Carbonell et al., 2021; Esper et al., 2023; Kläser et al., 2022). The Omicron variant, on the other hand, proved to be even more highly transmissible, but had the lowest hospitalization and mortality rates (Christensen et al., 2022). The observed differences in transmission and severity among COVID-19 variants is possibly related to the increased immunity among the infected people, either through vaccination or previous infection waves (Bhattacharyya & Hanage, 2022).

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 (Re3data.Org: GISAID, 2022a). Tracking of relative frequencies of variants from sequenced COVID-19 cases showed estimated time frames of the upsurge of specific variants: the ancestral strain was predominant, with 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 (Re3data.Org: GISAID, 2022b).

Machine learning is often used for health in the analysis of large datasets and the prediction of outcomes based on a variety of inputs including 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 several diseases including COVID-19, wherein they can give more than 90% accuracy in prediction and forecasting (Painuli et al., 2021). 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 (Mahdavi et al., 2021), and has been the topic of several researches (Hu et al., 2022; Mahdavi et al., 2021; Noy et al., 2022). 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 (Hu et al., 2022), support vector machines (Mahdavi et al., 2021), and decision tree ensembles (e.g., CatBoost, XGBoost, Random Forest) (Noy et al., 2022). A previous study utilized a publicly available national surveillance dataset to predict COVID-19 mortality in

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



the Philippines and identified age and history of hospital admission as significant predictors of disease outcome, but was limited to the ancestral strain which was present in the population at the start of the pandemic (Migriño & Batangan, 2021). This study aimed to describe COVID-19 outcomes by infection waves using machine learning.

METHODS

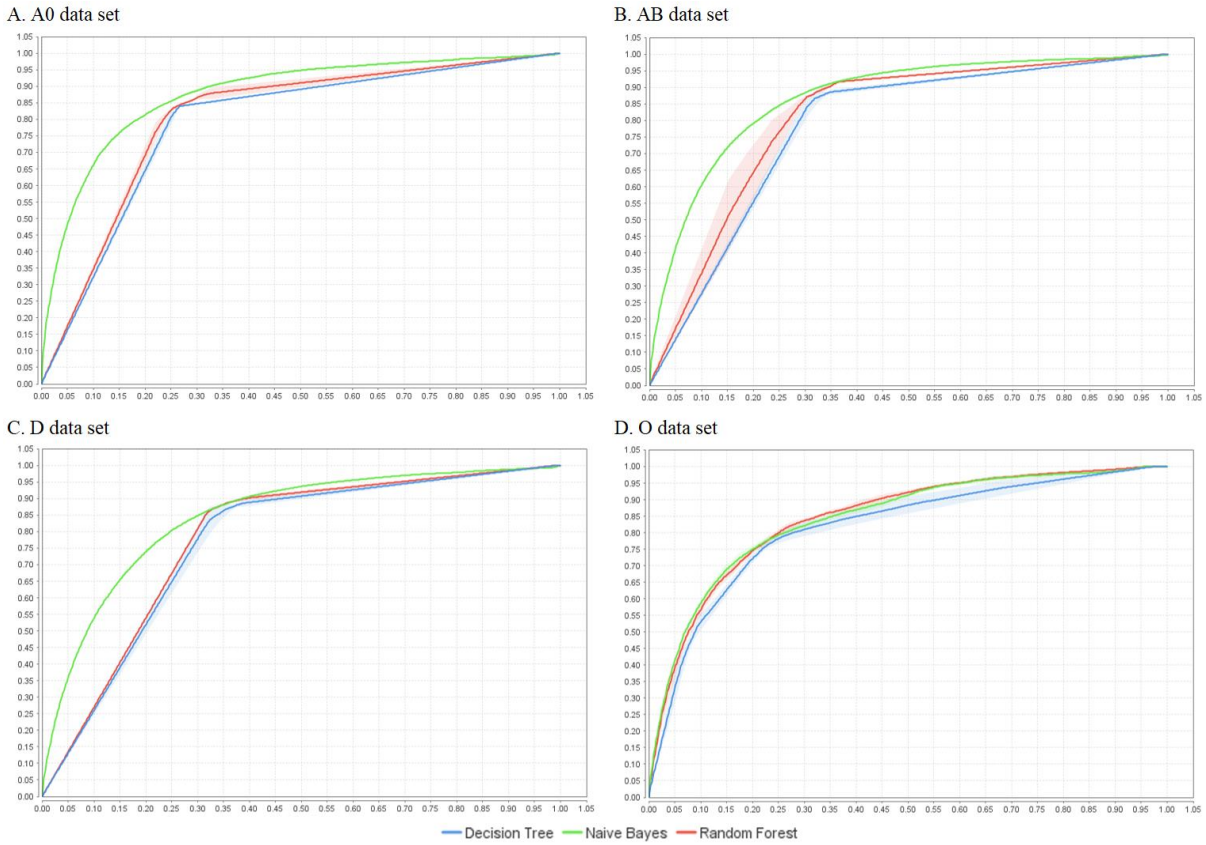
The study utilized a cross-sectional, documents review design using the DOH COVID Data Drop records as of September 24, 2022 (DOH, 2022). The database is a national record of all confirmed COVID-19 cases and was updated daily by the DOH Epidemiology Bureau. The full data set contained 3,934,777 cases and 22 attributes. Exploratory analysis of the raw data set was performed to visualize the reported cases and the different attributes. Ten attributes were included in the model generation which included Age, Sex, Admitted, RegionRes, ProvRes, CityMunRes, BarangayRes, Quarantined, Pregnanttab and RemovalType. Age_Group was generated to reclassify Age into nine bins based on the US CDC classification. DateRepConf, was retained only for splitting of the data sets (below). Cases with missing values for Age and RemovalType were dropped from the data set.

The four data sets are as follows: A0, with the predominant strain being ancestral, spans from January 30, 2020, to February 28, 2021; AB, where the Alpha and Beta variants were predominant, covers the period from March 1, 2021, to June 30, 2021; D, dominated by the Delta variant, ranges from July 1, 2021, to November 30, 2021; and O, with the Omicron variant as the predominant strain, extends from December 1, 2021, to September 24, 2022.

Descriptive statistics and statistical analysis, mainly t tests and Pearson's $\chi 2$ tests were generated with StataCorp 2013. Case fatality rates (CFR) were calculated using the equation 1.

Machine learning analyses were performed using RapidMiner Studio 9.10.008. The analysis is patterned after a similar study (Migriño & Batangan, 2021). Attribute selection was performed to screen out irrelevant attributes. Attribute selection was done individually using feature weights operators. RemovalType was used as the dependent variable in all data sets and had the value of RECOVERED or DIED. Hyperparameter grid optimizations were done by running fivefold cross-validation. Model generation was done using fivefold cross-validation using the optimized hyperparameters and RemovalType=DIED as the positive class set. For the training data sets, we conducted random undersampling (RUS) using simple random sampling to balance the RemovalType RECOVERED:DIED ratio. This training dataset was used to generate the decision tree models per fold. For the testing data sets, all recorded cases were used. The data sets were then running through naïve Bayes and random forest for comparison models. The

Figure 2
Reported COVID-19 cases in the Philippines by predominant variant



Note: [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.

receiver operating characteristic (ROC) curves of each model were generated. The performance metrics of each model were then extracted. These include area under the curve (AUC) and contingency table metrics such as accuracy, sensitivity/specificity and F-score.

The study was conducted at the San Beda University College of Medicine from August 2022 to April 2023. The research protocol was reviewed and approved under the study protocol SBU-RED 2022-020 by the San Beda University Research Ethics Board. We used the TRIPOD checklist for prediction model development as a framework for our methodology.

RESULTS

The final data set consisted of 3,896,206 cases, which comprised 99.02% of the raw data set reported cases, 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 final data set reported cases, respectively. Reported cases per day as well as the segmentation according to variants are visualized in f1. 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%). During 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 (See Table 1).

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

Characteristics	Overall (n=3,896,206)			Ancestral variant (n=571,905)			Alpha/Beta variant (n = 835,779)			Delta variant (n=1,414,651)			Omicron variant (n=1,073,871)		
	Recovered	Died	CFR	Recovered	Died	CFR	Recovered	Died	CFR	Recovered	Died	CFR	Recovered	Died	CFR
Overall Sex	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%
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%
Age															
Mean *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															
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															
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%	9,033	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%

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

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

Model	AUC		Accuracy		F-score		Sensitivity		Specificity	
A0 dataset										
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										
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										
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										
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%

Note: [a] Highlighted values are the largest values for each particular indicator across the three machine learning models

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 (9304, 7007,
4603 and 4323 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%) (See Table 1).

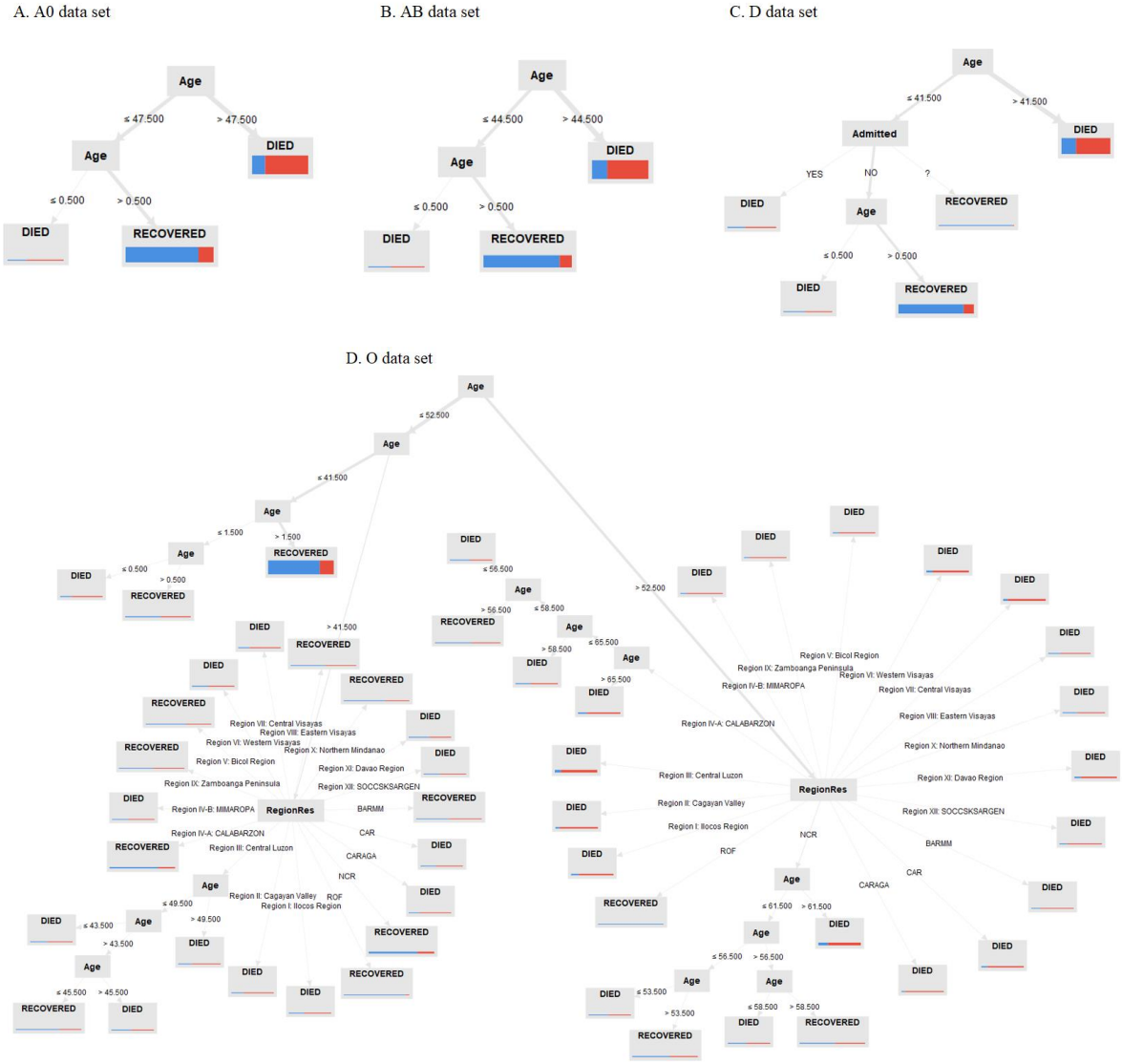
Out of the nine non-outcome attributes retained for model generation, only Age and Admitted were included in the models for data sets A0, AB and D. For the data set O, Age, Admitted and RegionRes were included in the model.

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 (See 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 (See Figure 2).

Figure 3 illustrated the decision tree models for the A0 and AB data sets were similar: they were composed of three levels, and each level (node) further splitting into two sub-levels (branches) (Figure 3A and Figure 3B, respectively). The D data set had four levels and had either two or three branches (Figure 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 (Figure 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 (Figure 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. The total and per-leaf number of cases, case outcomes, and other details of the decision tree models can be made available upon request.

Figure 3
Decision tree models of predicted outcomes from COVID-19 reported cases by data set[a]



Note: [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.

DISCUSSION

We generated four different decision tree models corresponding to the different predominant COVID-19 strain and variants in the Philippines, with age (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 (Admitted) as an additional attribute. The O data set generated a more complicated decision tree which incorporated age, admission history and region of residence (RegionRes) 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 (Venkatasubramaniam et al., 2017) and prediction of disease prognosis (Mahdavi et al., 2021; Migriño & Batangan, 2021).

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 R0 of 7 and 10, respectively, compared to 2.5 of the ancestral strain (Lorente-González et al., 2022; Rashedi et al., 2022). 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 immune evasion and reinfection regardless of vaccination status and previous infection (Mohsin & Mahmud, 2022; Wolter et al., 2022) due to enhanced viral replication efficiency in the bronchus (Hui et al., 2022).

Previous studies have found that the severity among the ancestral, Alpha and Delta variants are comparable (Esper et al., 2023; Hu et al., 2022), but the severity of the Omicron variant has consistently been lower (Esper et al., 2023; Lewnard et al., 2022; Wang et al., 2022; Wolter et al., 2022). This may be due to lower replication competence of the Omicron variant in the lung parenchyma (Hui et al., 2022). 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) (Bhopal & Bhopal, 2020; Migriño & Batangan, 2021) 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 CFR regardless of predominant COVID-19 variant. This general trend has been documented in previous studies (CDC, 2022; Endeshaw & Campbell, 2022; Esper et al., 2023; Malundo et al., 2022; SeyedAlinaghi et al., 2021). However, we noticed a pattern similar to a previous Philippine study on cases of the ancestral variant (Migriño & Batangan, 2021): 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 shows a generally increasing trend in CFR (CDC, 2022) but a study by Khera et al. (2021) supports our findings and attributed this "U-shaped" phenomenon to several factors such as children's differential expression of ACE-2 receptors, more robust innate immune system (except for newborns), and lesser exposure due to public health measures.

Decision trees as a machine learning algorithm offer models that are easily interpretable in both healthcare and policy settings and have thus been widely used in such fields (Sánchez-Montañés et al., 2020; Serrano, 2021), including in the analysis of COVID-19. This algorithm enables predictive analysis (such as mortality analysis) of large and even non-parametric data such as epidemiologic data (Venkatasubramaniam et al., 2017). 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 (Migriño & Batangan, 2021; Yadaw et al., 2020) 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 (Argosino, 2021; DOH, 2023). 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 (Malundo et al., 2022). Third, the D model incorporated history of admission as a splitting criterion, similar to a previous study (Migriño & Batangan, 2021). The previous hospitalization guidelines for COVID-19 patients in the Philippines prioritize admission of only severe and critical COVID-19 cases (DOH, 2020; Migriño & Batangan, 2021), 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 of the early COVID-19 ancestral wave in the Philippines (Migriño & Batangan, 2021) 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 (Carbonell et al., 2021; Pan et al., 2020; Talabis et al., 2021), length and severity of lockdowns, population or industrial

composition (Jiang et al., 2022; Talabis et al., 2021), and previous infection or vaccination rates (Bhattacharyya & Hanage, 2022; Kläser et al., 2022; Stein et al., 2023; Wang et al., 2022). Repatriated overseas Filipinos (ROF), on the other hand, are only allowed to return to the country if they are well enough to travel, hence the lower CFR among this cohort regardless of infection wave.

Surveillance data sets during the pandemic are often imbalanced in that the number of recoveries vastly outnumber reported deaths. We used under sampling techniques to control 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 (Serrano, 2021). We utilized similar techniques from a previous study (Migriño & Batangan, 2021) to reduce overfitting: removing irrelevant or highly correlated attributes during exploratory analysis, pre-pruning and pruning during training, and optimizing the hyperparameters for the highest sensitivity.

In the Islamic concept, maintaining health and preventing disease is part of everyone's responsibility to oneself, family and society. Age as the main predictor in reported COVID-19 cases shows how important it is to maintain health in every phase of life. The Prophet taught the importance of taking care of the body as a mandate from God. Therefore, self-protection through health education from young to old is a form of worship, in line with the Tirmidhi hadith No. 4977 which states

> *"The best gift to children from parents is their correct training".*

The findings in this study, where older age groups have higher mortality rates, underscore the need for special attention to the elderly, as taught in the Qur'an (QS. Al-Isra/17:23) which emphasizes respect and protection to parents. In the context of the pandemic, this could be translated as an imperative to ensure that the elderly get priority access to health care, vaccinations and other preventive measures. Furthermore, the pattern found in CFR by age, with young children and the elderly being more vulnerable, points to the importance of maintaining a balance between keeping children healthy and protecting the elderly. Children are considered a trust, and caring for them is a responsibility that is not only physical but also spiritual. All religions have directed humanity to take care of future generations, which can be applied in the form of health protection during a pandemic.

This 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. These important predictors of mortality include comorbidities, vaccination status and sociodemographic information. This may have negatively contributed to the performance metrics of our models, particularly the low accuracy and f-scores due to low class recall for mortality, which is an internal bias in surveillance datasets. 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 classification bias of reported cases, particularly those whose reported dates were near the boundaries of our infection wave timelines. Lastly, another possible source of bias is the classification bias of the surveillance dataset itself, in which classifying COVID-19 cases and deaths were based on unstandardized metrics especially during the early phase of the pandemic.

This study integrates the values of compassion, care for the sick and vulnerable, and the importance of safeguarding the sanctity of life. The use of technology, such as decision trees and other machine learning algorithms, in the timely identification of different risk profiles of different population groups during public health events enables healthcare workers and policy makers to more effectively fulfill their moral and societal imperative and obligations of service through protection of human life, promotion of the public's well-being, and instilling justice by allocating proportional resources to areas based on need.

CONCLUSIONS

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. Most cases in the Philippines occurred during the Delta and Omicron waves, and transmissibility was higher for the variants compared with the ancestral strain. The National Capital Region tallied the highest overall case rates, while Region VII recorded the highest case fatality rates which was observed during the Delta wave. The mortality pattern of the Omicron variant was significantly different from the preceding variants, consistent with studies from other countries. Our

decision tree models also reinforce the strong influence of increasing age in predicting COVID-19 outcomes regardless of SARS-CoV-2 variant. However, our Omicron model suggests possible differences in the impact of COVID-19 across different administrative regions in the country. Our study provides a simple framework in using machine learning to analyze publicly available surveillance data to monitor emerging or ongoing public health events such as outbreaks or epidemics. 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 incorporate relevant clinical factors such as presence of comorbidities, previous infections and vaccination status to provide a more comprehensive and robust analysis of mortality predictors. We also recommend relevant government agencies to enhance epidemiological data collection, analysis and dissemination to aid researchers and policymakers.

ACKNOWLEDGEMENT
The authors would like to thank the San Beda University Research and Development Center and the San Beda Office of the Vice President for Research and Innovation for the overall support to the study.

FUNDING
The study was funded in part by an operational grant from the San Beda University Office of Research and Innovation.

AUTHORS' CONTRIBUTIONS
Julius R. Migriño Jr. designed the study, formulated the concept, wrote and reviewed the manuscript, collected and acquired the data. Ani R. U. Batangan formulated the concept, wrote and reviewed the manuscript. Rizal M. R. Abello wrote and reviewed the manuscript. All Authors analyzed the data, revised manuscript, performed the filed work, and approved the final manuscript.

AUTHORS' INFORMATION
Julius R. Migriño Jr. is an associate Professor and Unit Head - Office for Medical Education in College of Medicine, San Beda University, Philippines. He is also an adjunct faculty, School of Medicine and Public Health, Ateneo de Manila University, Philippines. Ani Regina U. Batangan and Rizal Michael R. Abello is a researcher of part-time faculty in College of Medicine, San Beda University, Philippines.

COMPETING INTERESTS
The author(s) declare no potential conflict of interest with respect to the research, authorship, and/or publication of this article.

REFERENCES

Argosino, F. (2021, November 9). *COVID-19 response: A timeline of community quarantine, lockdowns, alert levels*. Manila Bulletin. https://mb.com.ph/2021/11/09/covid-19-response-a-timeline-of-community-quarantine-lockdowns-alert-levels/

Bhattacharyya, R. P., & Hanage, W. P. (2022). Challenges in Inferring Intrinsic Severity of the SARS-CoV-2 Omicron Variant. *New England Journal of Medicine*, *386*(7), e14. https://doi.org/10.1056/NEJMp2119682

Bhopal, S. S., & Bhopal, R. (2020). Sex differential in COVID-19 mortality varies markedly by age. *The Lancet*, *396*(10250), 532–533. https://doi.org/10.1016/S0140-6736(20)31748-7

Carbonell, R., Urgelés, S., Rodríguez, A., Bodí, M., Martín-Loeches, I., Solé-Violán, J., Díaz, E., Gómez, J., Trefler, S., Vallverdú, M., Murcia, J., Albaya, A., Loza, A., Socias, L., Ballesteros, J. C., Papiol, E., Viña, L., Sancho, S., Nieto, M., COVID-19 SEMICYUC Working Group. (2021). 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. *The Lancet Regional Health. Europe*, *11*, 100243. https://doi.org/10.1016/j.lanepe.2021.100243

CDC. (2022, December 28). *Risk for COVID-19 Infection, Hospitalization, and Death By Age Group*. Centers for Disease Control and Prevention. https://www.cdc.gov/coronavirus/2019-ncov/covid-data/investigations-discovery/hospitalization-death-by-age.html

Christensen, P. A., Olsen, R. J., Long, S. W., Snehal, R., Davis, J. J., Saavedra, M. O., Reppond, K., Shyer, M. N., Cambric, J., Gadd, R., Thakur, R. M., Batajoo, A., Mangham, R., Pena, S., Trinh, T., Kinskey, J. C., Williams, G., Olson, R., Gollihar, J., & Musser, J. M. (2022). 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*, *192*(4), 642–652. https://doi.org/10.1016/j.ajpath.2022.01.007

DOH. (2020). *Guidelines on the Risk-Based Public Health Standards for COVID-19 Mitigation* (Administrative Order No. 2020-0015). Department of Health. https://www.doh.gov.ph/sites/default/files/health-update/ao2020-0015.pdf

DOH. (2022, October 7). *COVID-19 Tracker | Department of Health website*. https://doh.gov.ph/covid19tracker

DOH. (2023, February 5). *Updates on COVID-19 Vaccines | COVID-19 Vaccination Dashboard*. https://doh.gov.ph/vaccines

Endeshaw, Y., & Campbell, K. (2022). Advanced age, comorbidity and the risk of mortality in COVID-19 infection. *Journal of the National Medical Association*, *114*(5), 512–517. https://doi.org/10.1016/j.jnma.2022.06.005

Esper, F. P., Adhikari, T. M., Tu, Z. J., Cheng, Y.-W., El-Haddad, K., Farkas, D. H., Bosler, D., Rhoads, D., Procop, G. W., Ko, J. S., Jehi, L., Li, J., & Rubin, B. P. (2023). Alpha to Omicron: Disease Severity and Clinical Outcomes of Major SARS-CoV-2 Variants. *The Journal of Infectious Diseases*, *227*(3), 344–352. https://doi.org/10.1093/infdis/jiac411

Hu, Z., Huang, X., Zhang, J., Fu, S., Ding, D., & Tao, Z. (2022). Differences in Clinical Characteristics Between Delta Variant and Wild-Type SARS-CoV-2 Infected Patients.

Frontiers in Medicine, *8*. https://www.frontiersin.org/articles/10.3389/fmed.2021.792135

Hui, K. P. Y., Ho, J. C. W., Cheung, M., Ng, K., Ching, R. H. H., Lai, K., Kam, T. T., Gu, H., Sit, K.-Y., Hsin, M. K. Y., Au, T. W. K., Poon, L. L. M., Peiris, M., Nicholls, J. M., & Chan, M. C. W. (2022). SARS-CoV-2 Omicron variant replication in human bronchus and lung ex vivo. *Nature*, *603*(7902), Article 7902. https://doi.org/10.1038/s41586-022-04479-6

Jiang, Y., Laranjo, J. R., & Thomas, M. (2022). COVID-19 lockdown policy and heterogeneous responses of urban mobility: Evidence from the Philippines. *PLOS ONE*, *17*(6), e0270555. https://doi.org/10.1371/journal.pone.0270555

Johnson, A. G., Amin, A., Ali, A., & et al. (2022). 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. Morbidity and Mortality Weekly Report*, *71*. https://doi.org/10.15585/mmwr.mm7104e2

Khera, N., Santesmasses, D., Kerepesi, C., & Gladyshev, V. N. (2021). COVID-19 mortality rate in children is U-shaped. *Aging (Albany NY)*, *13*(16), 19954–19962. https://doi.org/10.18632/aging.203442

Kläser, K., Molteni, E., Graham, M., Canas, L. S., Österdahl, M. F., Antonelli, M., Chen, L., Deng, J., Murray, B., Kerfoot, E., Wolf, J., May, A., Fox, B., Capdevila, J., Modat, M., Hammers, A., Spector, T. D., Steves, C. J., Sudre, C. H., Duncan, E. L. (2022). 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. *Scientific Reports*, *12*(1), Article 1. https://doi.org/10.1038/s41598-022-14016-0

Lewnard, J. A., Hong, V. X., Patel, M. M., Kahn, R., Lipsitch, M., & Tartof, S. Y. (2022). 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. *Nature Medicine*, *28*(9), Article 9. https://doi.org/10.1038/s41591-022-01887-z

Lorente-González, M., Suarez-Ortiz, M., & Landete, P. (2022). Evolution and Clinical Trend of SARS-CoV-2 Variants. *Open Respiratory Archives*, *4*(2), 100169. https://doi.org/10.1016/j.opresp.2022.100169

Mahdavi, M., Choubdar, H., Zabeh, E., Rieder, M., Safavi-Naeini, S., Jobbagy, Z., Ghorbani, A., Abedini, A., Kiani, A., Khanlarzadeh, V., Lashgari, R., & Kamrani, E. (2021). A machine learning based exploration of COVID-19 mortality risk. *PLOS ONE*, *16*(7), e0252384. https://doi.org/10.1371/journal.pone.0252384

Malundo, A. F. G., Abad, C. L. R., Salamat, M. S. S., Sandejas, J. C. M., Poblete, J. B., Planta, J. E. G., Morales, S. J. L., Gabunada, R. R. W., Evasan, A. L. M., Cañal, J. P. A., Santos, J. A., Manto, J. T., Mercado, M. E. P., Rojo, R. D., Ornos, E. D. B., & Alejandria, M. M. (2022). Predictors of mortality among inpatients with COVID-19 infection in a tertiary referral center in the Philippines. *IJID Regions*, *4*, 134–142. https://doi.org/10.1016/j.ijregi.2022.07.009

Migriño, J. R., & Batangan, A. R. U. (2021). Using machine learning to create a decision tree model to predict outcomes of COVID-19 cases in the Philippines. *Western Pacific Surveillance and Response*, *12*(3), Article 3. https://doi.org/10.5365/wpsar.2021.12.3.831

Mohsin, M., & Mahmud, S. (2022). Omicron SARS-CoV-2 variant of concern: A review on its transmissibility, immune evasion, reinfection, and severity. *Medicine*, *101*(19), e29165. https://doi.org/10.1097/MD.0000000000029165

Noy, O., Coster, D., Metzger, M., Atar, I., Shenhar-Tsarfaty, S., Berliner, S., Rahav, G., Rogowski, O., & Shamir, R. (2022). A machine learning model for predicting deterioration of COVID-19 inpatients. *Scientific Reports*, *12*(1), Article 1. https://doi.org/10.1038/s41598-022-05822-7

Painuli, D., Mishra, D., Bhardwaj, S., & Aggarwal, M. (2021). Forecast and prediction of COVID-19 using machine learning. *Data Science for COVID-19*, 381–397. https://doi.org/10.1016/B978-0-12-824536-1.00027-7

Pan, J., St. Pierre, J. M., Pickering, T. A., Demirjian, N. L., Fields, B. K. K., Desai, B., & Gholamrezanezhad, A. (2020). Coronavirus Disease 2019 (COVID-19): A Modeling Study of Factors Driving Variation in Case Fatality Rate by Country. *International Journal of Environmental Research and Public Health*, *17*(21), 8189. https://doi.org/10.3390/ijerph17218189

Rashedi, R., Samieefar, N., Akhlaghdoust, M., Mashhadi, M., Darzi, P., & Rezaei, N. (2022). Delta Variant: The New Challenge of COVID-19 Pandemic, an Overview of Epidemiological, Clinical, and Immune Characteristics. *Acta Bio Medica : Atenei Parmensis*, *93*(1), e2022179. https://doi.org/10.23750/abm.v93i1.12210

Re3data.Org: GISAID. (2022a, October 8). *Global COVID-19 submission tracker*. GISAID; re3data.org - Registry of Research Data Repositories. https://gisaid.org/submission-tracker-global/

Re3data.Org: GISAID. (2022b, October 8). *hCoV-19 Variants Dashboard*. GISAID; re3data.org - Registry of Research Data Repositories. https://gisaid.org/hcov-19-variants-dashboard/

Sánchez-Montañés M, Rodríguez-Belenguer P, Serrano-López AJ, Soria-Olivas E, Alakhdar-Mohmara Y. Machine learning for mortality analysis in patients with COVID-19. *Int J Environ Res Public Health*. 2020. November 12;17(22):8386. https://doi.org/10.3390/ijerph17228386

Serrano, L. G. (2021). *Grokking machine learning*. Manning Publications.

SeyedAlinaghi, S., Mirzapour, P., Dadras, O., Pashaei, Z., Karimi, A., MohsseniPour, M., Soleymanzadeh, M., Barzegary, A., Afsahi, A. M., Vahedi, F., Shamsabadi, A., Behnezhad, F., Saeidi, S., Mehraeen, E., & Shayesteh Jahanfar. (2021). Characterization of SARS-CoV-2 different variants and related morbidity and mortality: A systematic review. *European Journal of Medical Research*, *26*(1), 51. https://doi.org/10.1186/s40001-021-00524-8

Stein, C., Nassereldine, H., Sorensen, R. J. D., Amlag, J. O., Bisignano, C., Byrne, S., Castro, E., Coberly, K.,

Collins, J. K., Dalos, J., Daoud, F., Deen, A., Gakidou, E., Giles, J. R., Hulland, E. N., Huntley, B. M., Kinzel, K. E., Lozano, R., Mokdad, A. H., Lim, S. S. (2023). Past SARS-CoV-2 infection protection against re-infection: A systematic review and meta-analysis. *The Lancet*, *0*(0). https://doi.org/10.1016/S0140-6736(22)02465-5

Talabis, D. A. S., Babierra, A. L., Buhat, C. A. H., Lutero, D. S., Quindala, K. M., & Rabajante, J. F. (2021). Local government responses for COVID-19 management in the Philippines. *BMC Public Health*, *21*(1), Article 1. https://doi.org/10.1186/s12889-021-11746-0

Venkatasubramaniam, A., Wolfson, J., Mitchell, N., Barnes, T., JaKa, M., & French, S. (2017). Decision trees in epidemiological research. *Emerging Themes in Epidemiology*, *14*(1), 11. https://doi.org/10.1186/s12982-017-0064-4

Wang, C., Liu, B., Zhang, S., Huang, N., Zhao, T., Lu, Q., & Cui, F. (2022). 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. *Journal of Medical Virology*, 10.1002/jmv.28118. https://doi.org/10.1002/jmv.28118

WHO. (2022, October 4). *Tracking SARS-CoV-2 variants*. World Health Organization. https://www.who.int/activities/tracking-SARS-CoV-2-variants

WHO. (2023). *WHO Coronavirus Disease (COVID-19) Dashboard* [Dashboard]. WHO Coronavirus Disease (COVID-19) Dashboard. https://covid19.who.int

Wolter, N., Jassat, W., Walaza, S., Welch, R., Moultrie, H., Groome, M., Amoako, D. G., Everatt, J., Bhiman, J. N., Scheepers, C., Tebeila, N., Chiwandire, N., Plessis, M. du, Govender, N., Ismail, A., Glass, A., Mlisana, K., Stevens, W., Treurnicht, F. K., Cohen, C. (2022). Early assessment of the clinical severity of the SARS-CoV-2 omicron variant in South Africa: A data linkage study. *The Lancet*, *399*(10323), 437–446. https://doi.org/10.1016/S0140-6736(22)00017-4

Yadaw, A. S., Li, Y., Bose, S., Iyengar, R., Bunyavanich, S., & Pandey, G. (2020). Clinical features of COVID-19 mortality: Development and validation of a clinical prediction model. *The Lancet Digital Health*, *2*(10), e516–e525. https://doi.org/10.1016/S2589-7500(20)30217-X