Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Computers in Biology and Medicine 149 (2022) 105969

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Predicting COVID-19 disease severity from SARS-CoV-2 spike protein sequence by mixed effects machine learning

Bahrad A. Sokhansanj *, Gail L. Rosen

Ecological and Evolutionary Signal Processing & Informatics Laboratory, Drexel University, 3100 Chestnut St., Philadelphia, PA, 19104, United States of America

ARTICLE INFO

Keywords:
Viral genomics
COVID-19
SARS-CoV-2
Bioinformatics
Machine learning

ABSTRACT

Epidemiological studies show that COVID-19 variants-of-concern, like Delta and Omicron, pose different risks for severe disease, but they typically lack sequence-level information for the virus. Studies which do obtain viral genome sequences are generally limited in time, location, and population scope. Retrospective meta-analyses require time-consuming data extraction from heterogeneous formats and are limited to publicly available reports. Fortuitously, a subset of GISAID, the global SARS-CoV-2 sequence repository, includes "patient status" metadata that can indicate whether a sequence record is associated with mild or severe disease. While GISAID lacks data on comorbidities relevant to severity, such as obesity and chronic disease, it does include metadata for age and sex to use as additional attributes in modeling. With these caveats, previous efforts have demonstrated that genotype-patient status models can be fit to GISAID data, particularly when country-of-origin is used as an additional feature. But are these models robust and biologically meaningful? This paper shows that, in fact, temporal and geographic biases in sequences submitted to GISAID, as well as the evolving pandemic response, particularly reduction in severe disease due to vaccination, create complex issues for model development and interpretation. This paper poses a potential solution: efficient mixed effects machine learning using GPBoost, treating country as a random effect group. Training and validation using temporally split GISAID data and emerging Omicron variants demonstrates that GPBoost models are more predictive of the impact of spike protein mutations on patient outcomes than fixed effect XGBoost, LightGBM, random forests, and elastic net logistic regression models.

1. Introduction

Throughout the COVID-19 pandemic, SARS-CoV-2 has mutated in ways that have significantly impacted pathogenesis. Epidemiological studies can show different risks of severe disease due to different COVID-19 variants, such as Delta and Omicron, but typically lack resolution at the level of specific combinations of changes in viral genome sequences. The emergence of the COVID-19 pandemic, however, has coincided with the widespread availability of lower cost, rapid whole genome sequencing. As of writing, over 10 million SARS-CoV-2 sequences were available to researchers from the GISAID website (http://www.gisaid.org) [1,2]. GISAID includes a metadata field for "patient status" for a subset of sequences, which represents a potentially unparalleled resource for genetic analysis. If its potential can be unlocked, GISAID could provide the data necessary to develop a model of disease severity based on viral genotype and the limited patient characteristics available on GISAID, age and gender.

Clinical data which differentiate SARS-CoV-2 genotype generally do so at the level of lineages using the Pango nomenclature [3,4], or most commonly variants of concern (VOC) [5] based on those lineages, such as Alpha (Pango lineage designation B.1.1.7), Beta (B.1.351), Delta (B.1.167.2), and Omicron (BA.1 and BA.2). VOC designations generally refer not only to the original lineage but other "sublineages", such as AY.x sublineages of Delta and Omicron sublineages, such as BA2.12.1, B.4, and B.5. Studies have shown differences in case outcomes between lineages, supported in at least some cases with *in vitro* or animal model evidence for changes in virulence. The Delta variant had clear-cut increases in transmissibility and virulence, as indicated by both epidemiological estimates [6] and laboratory studies that show increased fitness over previous variants, including enhanced viral replication due to modification in the furin cleavage site of the spike protein [7,8]. Alpha, also appears to have resulted in more severe disease than the ancestral genome [9–12]. Overall, while Alpha resulted in elevated hospitalizations, ICU admissions, and other markers of severe outcomes as compared to ancestral lineages, Delta appears to have been yet more severe than Alpha [13–16]. By contrast, following Delta, Omicron

* Corresponding author.
E-mail addresses: bahrad@molhealtheng.com (B.A. Sokhansanj), glr26@drexel.edu (G.L. Rosen).
URL: http://www.drexeleesi.com (G.L. Rosen).

https://doi.org/10.1016/j.compbiomed.2022.105969
Received 6 June 2022; Received in revised form 11 July 2022; Accepted 13 August 2022
Available online 17 August 2022

has appeared to result in less severe disease, even when controlling for vaccination [17]. Lower risks of hospitalization, and in particular shorter hospital stays, reduced ICU admissions, and less use of ventilation rates, particularly as compared to Delta, have been shown in studies from Denmark [18], the United States, [19,20], and the United Kingdom [21,22]. Epidemiological data for Omicron are consistent with *in vitro* and animal studies, which have shown a reduction in lower lung infectivity, deficient cell entry, and a reduction in syncytium formation due to reduced ability of the spike protein to mediate plasma membrane fusion [23–26].

The aforementioned differences in case outcome are between apparently sequential emergent SARS-CoV-2 genetic lineages. In fact, however, changes to SARS-CoV-2 properties often implicate combinations of multiple mutations that emerge simultaneously—and then sometimes revert in whole or in part as the virus continues to evolve [27,28]. For example, Delta, originally Pango-designated B.1.617.2, has spawned complex sublineages (generally identified by an AY prefix) with distinct immune evasion and virulence properties—and can genetically share more in common with other lineages than the one from which they apparently branched [29,30]. During a long-term infection, a spike protein may emerge with multiple variations, i.e., a "long branch" divergence from the phylogenetic tree, a process hypothesized as the origin of the Omicron variant of concern (initially identified as Pango lineage B.1.592, and soon redesignated as BA.1 and BA.2) [31]. As Omicron has become dominant, subsequent Omicron subvariants have emerged, including subvariants of BA.1 and BA.2, as well as BA.3, BA.4, and BA.5 [32]. These Omicron subvariants are characterized by apparent recombination of mutations, as well as the appearance of mutations that look similar to those in previous variants, such as Alpha [33]. Recombinant variants of Delta and Omicron have also been identified in larger numbers; these are designated in the Pango scheme with an initial "X", e.g., XD and XE [34].

Conventional techniques for studying the effect of SARS-CoV-2 genotype on COVID-19 mortality and symptom severity have included analysis of single nucleotide polymorphisms (SNP) and genome wide association studies [35–37], literature meta-analysis [38]. Individual studies have been limited in the number of patients, and meta-analyses are time-consuming, complicated by having to access underlying data, and generally exclude unpublished data. The viral sequences from many such published and unpublished studies have been deposited in the GISAID global SARS-CoV-2 sequence repository, along with data for patient status. Various groups have investigated the incidence and prevalence of mutations in GISAID entries with clinical metadata [39,40]. While these studies have yielded some potential candidate mutations, the data are often conflicting or not necessarily consistent with epidemiological or laboratory observations. For example, some of the latter studies showed a link between the D614G mutation, which emerged early on as cases spread from Asia to Europe and North America, with increased disease burden. But another study of case fatality rates by region did not find a correlation with the dominant clade in that region [41].

In general, efforts to build logistic regression and other statistical models to predict mild versus severe disease on GISAID data have shown that much of the explanatory power is provided by patient age, gender, and region of origin, rather than clade or lineage [42]. However, it has been shown that adding sequence data to a logistic regression method can produce a more accurate prediction of severe versus mild disease than one with only age, gender, and region, although the difference was not particularly large [43]. Moreover, an updated of the latter model trained and tested on more recent sequence data resulted in deteriorated performance, even when employing a more accurate random forests classifier method [44]. Another group employed a powerful gradient-boosted decision tree ensemble classifier method, XGBoost, and found that models evaluated using temporally split data, i.e. trained on earlier sequences and tested on later-emerging sequences, substantially outperformed models evaluated using cross-validation, in which the training and test samples are randomly selected [45]. The authors analyzed the trained models to identify mutations associated with increased severe disease risk. However, key findings such as the V1176F mutation, while present in VOC, have not been specifically linked to disease severity in laboratory or epidemiological studies. Other methods have been employed, including deep neural networks [46,47] and Bayesian multinomial logistic regression to infer growth rate, and thus viral fitness, from individual sequence mutations [48]. However, there is no consensus modeling method for analyzing GISAID data, and the complexity of the data has not been fully analyzed.

The modeling approach in this paper begins by first evaluating the trends and structure of GISAID data—a critical step for developing robust genotype-phenotype models. There are two issues that impact the use of GISAID as a data source: (1) The nature of the pandemic has changed over time, with more screening of asymptomatic or mild cases, as well as improvements in therapeutics and widespread vaccination. (2) While unprecedented numbers of SARS-CoV-2 sequences have been deposited, that still represents only a sampling of viral infections worldwide. For example, as of April 2022, over half of all sequences in GISAID are from the United States and United Kingdom [49]. The set of sequences with GISAID patient metadata which we were able to curate (excluding illegible or unknown metadata fields) are an even smaller subsample. In practice, other work has shown that models trained on earlier records do not perform well on later sequence records [44,45]. In part due to the evolution of novel mutations, but it may also have to do with changes in the nature of what kind of sequences are submitted to GISAID over time. For example, in our previous work, we showed that, through September 2021, there had been a consistent increase in "mild" cases observed in the database [47]. In the analysis shown here, we identify that the latter temporal trends may have now stabilized, but that heterogeneity in the geographic origin of samples may be an important confounder.

Notably, other potential risk factors, such as obesity or chronic disease, are not provided in GISAID metadata. Moreover, while vaccination status is a metadata field, only very few samples include an entry for it. As a result, modeling efforts based on GISAID data will always lack information on known co-founders. Even so, GISAID does have many more samples than targeted studies that do include information about comorbidities, which at least mitigates potential data bias. Also, training om data after vaccination is more widespread, as shown here, can mitigate biases due to vaccination status. That said, the foregoing caveats are important for the work presented here as well as all efforts to model the effect of viral genotype on disease severity.

Taking the aforementioned observations and caveats into account, in this paper we examine the overall data set. Then, we identify a timeframe for model training that can result in more robust models for evaluation, and hypothesize that including geographic origin through the "country" metadata field in a mixed effects model will result in more robust models as well. We propose to use a recently-developed mixed effects machine learning method, GPBoost, which incorporates decision tree-boosting to efficiently train on large data sets with many features [50]. GPBoost is compared to conventional methods, including logistic regression, as well as ensemble decision-tree based methods, Random Forests [51], XGBoost [52], and LightGBM [53]. The best-performing methods, GPBoost, XGBoost, and LightGBM, are interpreted using SHAP (SHapley Additive exPlanations) [54], which has previously been used to interpret XGBoost models [45]. The modeling methodologies are evaluated on the spike protein sequence, as it binds to host cell receptors, mediates cell entry, is a key target for the immune response, and has a high rate of mutation [55–58]. Analyzing only the spike protein sequence further reduces the risk of overfitting and make models more computationally tractable.

2. Methods

2.1. Spike protein sequence collection and pre-processing

Spike protein sequences are obtained from a FASTA file available from the GISAID database (http://www.gisaid.org). The data for this study were downloaded on sequences that were submitted to and processed by GISAID as of April 15, 2022. Based on the metadata for collection date, the latest-collected sample in this data set was from April 10, 2022. GISAID performs various data curation tasks; of relevance here, Spike protein sequences are preprocessed by GISAID by multiple sequencing alignment, identifying ORFs, and translating nucleotide sequences to obtain protein FASTA files [2]. The FASTA file is parsed to obtain only those sequences for which patient metadata are available. (The section below details how patient metadata are obtained). The acknowledgment table for the sequences used in this study may be found at https://doi.org/10.55876/gis8.220606hk.

Many of the spike sequences are truncated due to sequencing gaps and errors. Therefore, the Spike protein sequences from the FASTA file are aligned with respect to the consensus Spike reference sequence (Wuhan-Hu-1 isolate) obtained by multiple sequence alignment of early genome sequences [59]. The alignments are generated using the local pairwise Striped Smith–Waterman (SSW) method [60,61], with BLOSUM62, implemented with the scikit-bio package in Python 3.8 [62]. Aligned sequences shorter than the reference (1273 residues) are front and/or end padded with a "*", and otherwise all indels are at positions corresponding to the reference. To preserve as many samples as possible, there is no filtering sequences with "*" (mask) or "X" (ambiguous amino acid).

2.2. Patient status metadata collection and pre-processing

The GISAID database provides an option to identify sequence records that include "patient metadata" and to download the metadata file with that information. This study includes the data from the records available for sequences collected by April 15, 2020: 414,297 records in total. (By comparison, at that time, the aforementioned Spike protein sequence FASTA file, that are from studies with and without metadata, included over 5 million sequences.) After metadata exclusions are applied (described in the following paragraphs), 163,496 samples remain available for machine learning. These records include an entry for "patient status" as well as metadata fields generally available for all SARS-CoV-2 sequences on GISAID, which include *inter alia* host, the continent/country/region of collection, Pango nomenclature lineage, NextStrain clade, sample collection and submission date, patient age, and patient gender. As an initial matter, all samples for which the host is not identified as "Human" are removed from the dataset.

The patient metadata consists of a single field with text provided by the submitter of the sequence. There are many different kinds of entries, including misspellings. As a preliminary step, these metadata entries are translated to a "Status". The "Status" translates different entries which may consist of different spellings or synonyms for the same activity, such as cases obtained by screening asymptomatic carriers. The table includes examples of entries assigned to these categories. Table 1 shows all of the unique metadata entries in the full patient metadata set (414,297 records) along with the corresponding "Status" designation. The resulting status is then categorized, generally following the commonly used case classification such as those defined by the United States National Institutes for Health (NIH) COVID-19 guidelines [63]. 1 shows the categories and the status designation. For example, sequences with metadata indicating ICU admission or mechanical ventilation are categorized as "Severe". Some metadata entries are categorized as "Unknown" even if they are not explicitly entered as such, as they do not contain information about the patient's status, for example some appear to refer to the age of the patient or to the location where the sample was taken. Metadata entries of "recovered" were also placed in the unknown category, as there was no indication of the severity of prior illness. Notably, the "Asymptomatic" category is defined to also include paucisymptomatic cases which are not expressly defined as "Mild". As such, there will be some overlap between those two categories.

The categories are then assigned to "Mild", "Severe", and "Unknown" classes, according to the NIH categories where there is sufficient information. 1 shows the class assignments for each category. For example, it is not clear whether "Alive" indices alive, but in an ICU, or alive and with mild symptoms. Accordingly, the "Alive" category is assigned to "Unknown". Similarly, a "Symptomatic" or "Alive" patient may have severe symptoms or have been hospitalized; therefore, the "Symptomatic" category is thus assigned to "Unknown". The "Released" category indicates release from prior hospitalization, and, therefore, "Released" is classified the same as "Hospitalized". Cases in the "Screening" category are classified as "Mild". These are cases with metadata entries such as "random screening", "community screening", and "airport screening;" as such, they are assumed to be from asymptomatic, or, at minimum, ambulatory individuals who were not hospitalized for COVID-19 symptoms at the time of sequencing. Cases in the Unknown class are dropped from the analysis.

The models described in this study also utilize metadata fields for "age" and "gender". Notably, the "gender" field includes entries that suggest it is being used interchangeably for gender and sex. For the purpose of simplicity and to align with the GISAID field names, the term "gender" is used in this paper. With respect to the gender field, any entry that is cognizable as Male or Female (e.g., misspellings, foreign language words such as "Homme", which is French for "man", etc.) are classified accordingly. Any other entry is classified as "Unknown" and excluded from analysis. The "age" metadata entries are assigned to an integer age where possible. Where the "age" entry is provided as a range, e.g., "'21-30", it is assigned to the mid-point, e.g., 25. Where the "age" field entry is "unknown" or a value that cannot be translated to an integer, the sample is excluded.

2.3. Machine learning

Five machine learning methods are used: (1) logistic regression with elastic net regularization [64] (referred to as "elastic net" or "logistic regression" in this paper), which has previously been utilized for genetic association studies [65,66]; (2) the random forests (RF) ensemble tree-based method [51], which is used to classify SARS-CoV-2 sequences to Pango lineages [3]; (3) eXtreme Gradient Boosting (XGBoost) [52], a decision tree-based ensemble learning method which has been used for SARS-CoV-2 nucleotide sequence classification [45] and which our group and others have previously used to classify protein sequences [67,68]; (4) LightGBM, which is a gradient-boosting method developed by Microsoft that grows trees leaf-wise, unlike XGBoost, which grows trees depth-wise, thus running much more efficiently while achieving comparable results [53,69]; and (5) GPBoost, which trains a mixed effects model including both features, implemented using decision trees (trained using LightGBM), and random effects at the group level [50].

2.3.1. Mixed effects models

Linear mixed effects models have been used for analyzing genetic studies where there are group-level random effects, such as in longitudinal studies and other sampling studies where there may be batch effects due to different laboratory methods being used for samples taken at different locations or times [70–73]. Eq. (1) shows the general matrix formulation for a mixed effects model.

$$y = F(X) + Zb + \epsilon \tag{1}$$

$F(X)$ is the row-size evaluation of function F, and ϵ. X and Z are fixed effects and random effects predictor variable matrices respectively, i.e., the rows of X are predictor variables for n observations (columns

of X). In this study, the random effects vector b is assumed to contain grouped (i.e. clustered) random effects. In this case, the columns of Z will be one-hot encoded (i.e., Z will be an incidence matrix with 1s and 0s) with the categorical variables that define the structure of groups. Assuming that the fixed effects model is linear, then Eq. (1) may be written in terms of groups as shown in Eq. (2).

$$yi = X_i\beta + Z_i b_i + \epsilon \tag{2}$$

In Eq. (2), the linear model for $F(X)$ is given as X_i is the $n_i \times p$ model matrix for fixed effects for observations in the ith group, where there are n features, β is a $p\times 1$ vector of model coefficients, and ϵ_i is the $n_i \times 1$ vector of errors for the ith group. Z_i is now a $n_i \times q$ model matrix of random effects for the ith group, where b_i is the $q\times 1$ vector of random effect coefficients for the ith group.

In this paper, groups of random effects are identified by country metadata. The means of b and ϵ, an unknown vector of random errors, are 0; accordingly, we take the mean of the model response in order to evaluate its predictions. We implement mixed effects machine learning with GPBoost, which has been made publicly available at https://github.com/fabsig/GPBoost. GPBoost is a highly efficient package for fitting mixed effects models to data, as it utilizes LightGBM tree-boosting to model fixed effects [50]. Further elaboration of the mathematical foundation of mixed effects models relevant to this paper can be found in [50]. GPBoost is thus able to handle the large feature set required to include the full spike protein sequence.

2.3.2. Feature representation

The input for machine learning are features vectors of integers for each sample, and training labels set at 0 (for Mild) and 1 (for Severe). Features are obtained as follows. After the alignment procedure described above, all of the resulting sequences have 1273 characters (amino acids, deletions, or masks). The sequences are tokenized, converting each character, including the deletion symbol "-", to a distinct nonzero integer. A position with padding mask "*" or ambiguous amino acids represented as X, B, J, or Z are considered to be missing data. Accordingly, they are a value of NAN for XGBoost, LightGBM, and GPBoost, which can then treat them as missing values; or, they are assigned a value of 0 for logistic regression and Random Forests, which cannot handle missing data. The age is represented as an integer, as describe above, and gender is treated as 0 and 1. In total, then, there are 1275 features: 1273 amino acid positions, age, and gender. As described in the paper, we also tested using the metadata for "Country" of origin of a case as a feature (increasing the number of features to 1276), or in the case of GPBoost, as a grouping of random effects in a mixed effects model. In that case, the "Country" was tokenized and represented as an integer using scikit-learn.

2.3.3. Model interpretation to obtain feature importance

The feature significance shown in the Results section for XGBoost, LightGBM, and GPBoost were obtained from SHAP (Shapley Additive eXplanations) values of terms for the test data set using the TreeExplainer method within the SHAP module (https://shap.readthedocs.io/) in Python 3.7 [54]. Among the principal reasons for selecting GPBoost to implement mixed effects machine learning was its compatibility with SHAP for interpretation [50,74]. Feature importance can also be derived for the aforementioned ensemble decision tree methods by computing, for example, the number of times a feature is used to split trees, or the gain in score towards the objective function obtained by splitting trees based on a feature [75,76]. However, we found no substantive differences between the features identified as significant using SHAP and those computed based on decision tree characteristics; moreover, SHAP not only estimates feature significance, but can also estimate whether a feature value tends to result in one classification or another.

2.3.4. Hyperparameter tuning and model implementation

For the results of this paper, training and testing data splits were determined by sample collection date as described in the Results section. Hyperparameter tuning was performed using a data set consisting of 60,196 samples collected between May 6, 2021 through November 2, 2021. Five-fold cross-validation was used to define training and testing splits, and the mean class prediction accuracy on the testing sets across three runs of the algorithm was computed for each hyperparameter combination. The hyperparameter combination with the highest accuracy was selected for the data presented in the Results. Other hyperparameter combinations were tested on that data and were not found to perform better than those that were used. The hyperparameters for the respective methods are as follows. Where not provided here, hyperparameters were set at their default values.

- **GPBoost.** The number of boosters was set at 2000 (values from 500 through 3000 were tested), maximum tree depth set to 30 (values from 10 to 50, as well as unlimited, were tested), maximum number of leaves set to 20 (tested 10 to 50), and learning rate set to 0.01 (tested 0.001 to 0.1).
- **LightGBM.** The number of boosters was set at 2000 (tested 500 to 3000), maximum tree depth set to 30 (tested 10 to 50 and unlimited), maximum number of leaves set to 20 (tested 10 to 50), and learning rate set to 0.01 (tested 0.001 to 0.1).
- **XGBoost.** The number of estimators was set at 2000 (tested 500 to 3000), maximum tree depth set to 20 (tested 10 to 50 and unlimited), lambda regularization set at 2.0 (tested 0.0 to 3.0), gamma set to 1.0 (tested 0.0 to 2.0), and learning rate set to 0.01 (tested 0.001 to 0.1).
- **Elastic Net.** The l1 ratio is set at 0.65 (tested 0.4 to 1.0), C is set to 0.1 (tested 0.01 to 0.8), and the maximum number of iterations was set to 1000 (tested 200 to 2000).
- **Random Forests.** The number of estimators is set at 500 (tested 200 to 2000), the maximum depth is set at unlimited (tested 10–50 and unlimited), the minimum number of samples required in a leaf node is set at 1 (tested 1–3), and the minimum number of samples required to split an internal node is set to 2 (tested 1–5).

The results in this paper were obtained using Python 3.7.13 or 3.94, `scikit-learn` package version 1.0.2 [62] (for elastic net and random forests methods), and the Python implementations for `xgboost` version 1.6.1, `gpboost` version 0.7.6.2, `lightgbm` version 2.2.3, and `shap` version 0.40.0. Training and hyperparameter tuning were performed on the Drexel University Research Computing Facility's Picotte high performance cluster using multithreaded implementations of the methods on Dell PowerEdge R640 servers with Intel® Xeon® Platinum 8268 CPUs. Model evaluation and visualization were performed in the Google Colab environment.

2.4. Resource availability

2.4.1. Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Dr. Bahrad A. Sokhansanj (bahrad@molhealtheng.com).

2.4.2. Data and code availability

- The datasets analyzed for this study were downloaded from GISAID EpiCoV database pursuant to the GISAID terms of use. They are availabile for download to users who register with GISAID at the website http://wwww.gisaid.org. The list of GISAID accession numbers used for this paper and data acknowledgments are available at https://epicov.org/epi3/epi_set/EPI_SET_20220606hk or https://doi.org/10.55876/gis8.220606hk.
- The code used for pre-processing and analysis in this paper has been deposited to and made publicly available from the authors' GitHub repository, https://github.com/EESI/covid_severity.

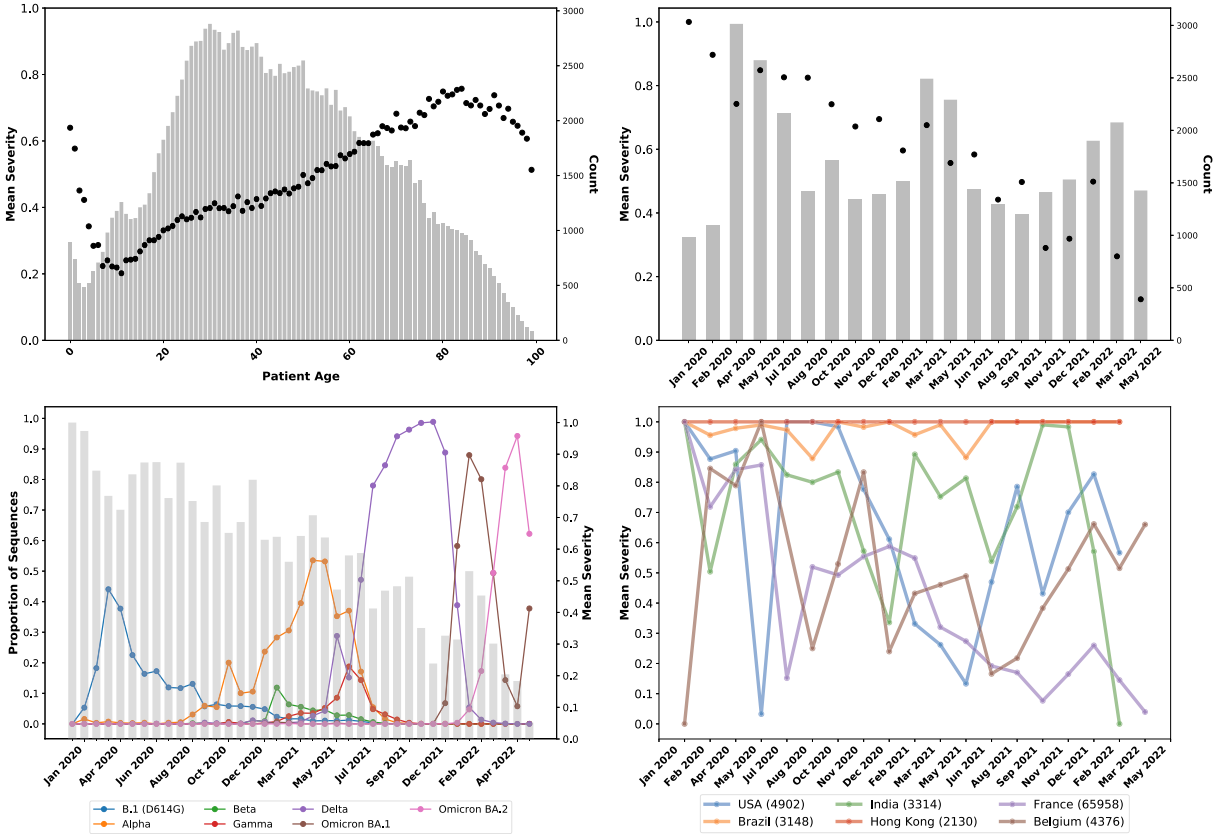

Fig. 1. Overview of age, sample collection date, and country metadata trends in GISAID data. (A – Upper Left) Mean case severity, where 0 is Mild and 1 is Severe, which equates to the probability of a severe case) by patient age in the GISAID database. The bars show the count of samples for each age. Increasing age trends with increasing severity, as expected, with differences at extremely low and old ages characterized by low sample counts. **(B – Upper Right)** Mean clinical severity (probability of severe case) by sample collection date recorded in the GISAID data. For clarity, data have been binned over time periods; the bars indicate the number of samples. Over time, the proportion of severe cases has declined, although that trend has been less consistent since Fall 2021. **(C – Lower Left)** Proportion of sequences in the GISAID patient data set (sequences with patient metadata) for principal variants, including B.1 (the ancestral lineage with the D614G which emerged in Northern Italy and New York in February–March 2020) and its sublineages, Alpha (B.1.1.7 and "Q" sublineages), Beta (B.1.351), Gamma (P.1), Delta (B.1.617.2 and AY sublineages), and the two major Omicron lineages, BA.1 and BA.2 (and their sublineages). The bars indicate the mean case severity for each date bin. The trends of sequential lineage waves in GISAID patient data appear to be consistent with the larger GISAID data set, i.e., showing successive Alpha, Delta, and Omicron waves. **(D – Lower Right)** Mean case severity of samples separated by GISAID metadata for the country where the sequence was collected for selected countries. The total number of sequences in the GISAID patient data set per country is shown within parentheses in the legend. Fluctuations in severity observed in countries appear due to systemic issues or differences in where samples are collected (e.g., in hospitals or outside settings) at different times.

- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

3. Results

3.1. Descriptive analysis of GISAID patient data

To develop an understanding of the GISAID patient data set (i.e. data with metadata subject to the exclusions described in the Methods section), we analyze the trends for severity for the metadata fields available in the GISAID data set: age, gender, sample collection date, and geographic origin of the case. To quantitatively measure severity trends, the classifications for Mild disease is assigned a severity level of 0, and Severe, a severity of 1. (The classifications are derived from patient status metadata based on Supplementary Tables S1 and S2 as described in the Methods.) Thereby, the mean of the severity values can be computed, which equates to the proportion of samples which are classified as Severe cases.

Fig. 1A shows the mean severity for each age from 0 to 100, as well as the number of samples in the GISAID patient data set for each age. In general, the fraction of severe cases increases with age, which has been a consistent feature of the pandemic [77,78]. Trends are different at the extremes, very young and old ages. Notably, there are far fewer cases at these ages, thus small biases in sample collection can have significant effects. For example, very young patients may be likelier to be observed in a hospital setting than observed in a screening study. Male sex has also been identified as a potential risk factor for more severe outcomes, such as ICU admission and death [79,80]. GISAID provides sex information in a "gender" metadata field. As shown in Fig. 2A, the relationship between increased age and increased proportion of severe cases are consistent throughout the pandemic.

Fig. 1B shows that, in addition to known risk factors, over time the mean case severity significantly decreases. The declining severity trend through 2020 in GISAID data is consistent with a period of improved COVID-19 therapeutics. For example, a Canadian study measured a decrease in case fatality rate (CFR) between the first and second waves prior to any vaccination, even when controlling for age and increased testing [81]. Later reduction is consistent with increased levels of COVID-19 vaccination reducing severe outcomes [82–85], as well as continued improvements in therapeutics such as monoclonal antibodies [86]. Notably, while the trend shows an overall decrease, which we had previously observed through October 2021, it is not monotonic, which an increase shown in late 2021 and early 2022. Moreover, the absolute level of severe cases suggests that while the trend of decreasing severity is consistent with the global trend of decreasing overall severity, the nature of reported cases also affects the trend. For example, in the initial first binned time periods, there

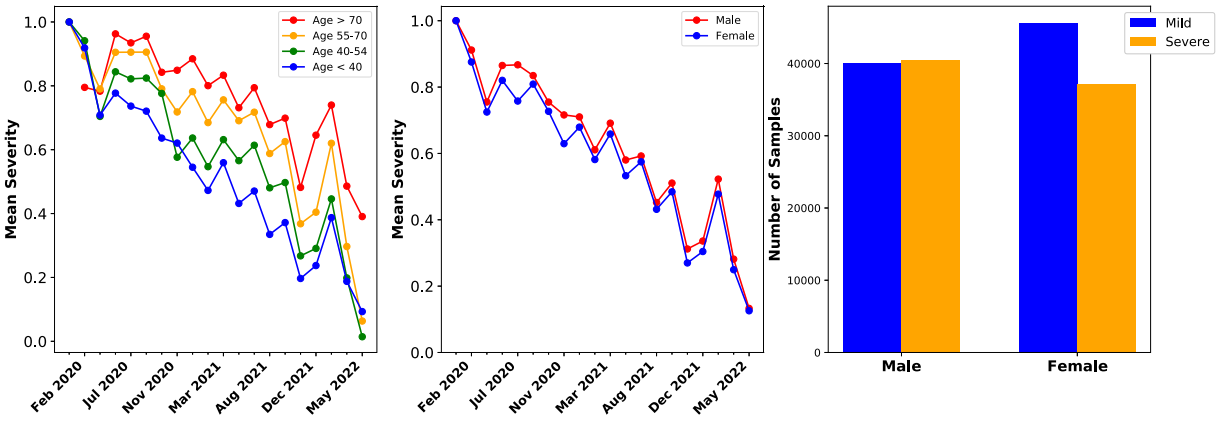

Fig. 2. Patient age and gender metadata trends in GISAID data. (A – Left) Mean clinical severity over time for patients in different age groups, showing that the overall trends are generally consistent across age groups, with older patients having mean severity as shown in Panel A. **(B – Middle)** Mean clinical severity, separating male and female samples, showing consistent trends across gender with male patients generally having a somewhat higher ratio of severe cases. **(C - Right)** Number of mild and severe cases across all samples split by gender, showing that there are more mild cases than severe among samples from female patients.

are over 70% severe cases. As illustrated in Fig. 3, the large majority of these are hospitalizations, although approximately 10% of samples are from dead patients according to the metadata entries. Studies of the initial pandemic waves in March–April 2020 did show that a CFR that approached or exceeded 10% [87–89]. Subsequent analysis estimated that the infection fatality rate (IFR) was likely much closer to 1%–2%, with the elevated CFR being due to underreporting of cases [88,89]. Among the passengers of the *Diamond Princess* cruise ship in February 2020, who were all tested, the IFR was 1% (in a population which skewed older). Accordingly, while the proportion of deaths in the GISAID patient data set did reflect community observations, at least in this timeframe, they were elevated due to so many cases being missed. However, the fraction of sequences submitted from dead patients continued to run at 4%–5% even through October 2021. We do observe that many cases have a metadata entry of "alive" or "live", as a contrast with "dead". But "alive" or "live" metadata cannot be identified as either Mild and Severe, and are thus excluded from the data set analyzed in this paper and shown in Fig. 1. Sequences from dead patients are thus overrepresented.

Fig. 2B and C show that GISAID gender metadata similarly indicate elevated severe disease among male patients. Samples with Male gender metadata are classified as 49.8% Mild and 50.2% Severe; by comparison, samples with Female gender metadata are 55.1% Mild and 44.9% Severe. An increased proportion of severe disease for older and male patients is consistently observed in GISAID samples collected at different dates over time. Notably, the difference in severity between male and female patients (defined according to gender metadata) was much greater in samples collected up to mid-2021, and has decreased since then. It is unclear whether this reflects a broader trend or is an artifact of where GISAID samples are collected.

Moreover, the number of hospitalizations is much more elevated even than the inflated hospitalization rates observed during the period of significant underreporting in early 2020. As Fig. 3 shows, even by March 2021, over 50% of GISAID samples being collected were from individuals who were either hospitalized or released from hospital, per their patient status metadata. This makes intuitive sense, since sequence samples with clinical information may well come from clinical settings, particularly hospitals. As a result, even though there is a steady increase in cases classified as Mild (see Fig. 3), this is likely at least in part because of a change in settings where sequences with metadata are collected, with more of a mix of outpatient settings. We observed a pronounced spike in the proportion of sequences annotated as Asymptomatic or collected from population screening studies in April–May 2021, from around 3% to 12%, which is consistent with changes in sampling sources.

Fig. 1C shows how the aforementioned distortions in the data can practically impact the development of genotype-severity models. There is a lower observed mean severity during the Delta wave as compared to the timeframe of Delta's emergence (when other lineages have a significant fraction of samples being collected) and before then, when Alpha was a plurality lineage. While this potentially could be due to continued vaccination resulting in less severe cases overall, and in turn fewer sequences from severe cases in the GISAID patient data set, Fig. 1C shows the trend of reduced severity reverses in the initial Omicron (BA.1) wave. However, vaccination and improved therapeutics, while certainly having resulted in a reduction in severe disease outcomes in general, do not explain the observed reduction of mean severity within the GISAID data set. We can show that as follows: Time-dependent changes in external conditions, such as increased vaccination rates, can be controlled for by looking only at the short timeframe where Alpha and Delta were collected in similar numbers, *c*irca May–June 2021. As Fig. 1C shows, the reduction in severity over time at the Alpha–Delta transition point is abrupt. Significantly, during May–June 2021, Delta samples were 60.7% Mild (5199 total samples) and Alpha samples were 52.2% Mild (8658 total samples). As a result, a model based on the GISAID data set will show that Delta is milder than Alpha, contradicting the epidemiological and laboratory evidence discussed above [13–16]. It is also not the case that Delta samples during the May-June 2021 timeframe were collected in countries with higher vaccination rates than Alpha samples. The main source of samples during this timeframe for both lineages was France (44% and 42% of Delta and Alpha respectively), and the second largest sources was Mexico (19% and 11%). Notably, in samples from France, Delta was 92.6% Mild and Alpha was 62.6% Mild, while in samples from Mexico, Delta was 27.1% Mild and Alpha was 38.4% Mild. Therefore, the observed decrease in severity from Alpha to Delta, due to this apparent artifact found in data from France, will confound a genotype-patient status model. Mutations associated with Delta will appear to result in reduced severity, in contradiction to epidemiological and other evidence.

Previous efforts in modeling GISAID patient data have found that including the location metadata for the country of origin of the sequence results in a more accurately predictive model [43]. As an initial matter, country and sequence will likely have some correlation, since mutations, including both the lineage an sublineage level, cluster between countries [90,91]. However, there is likely some variance in patient outcomes between countries due to differences in the enforcement of non-pharmaceutical interventions (NPI) which may be more protective of vulnerable populations, differences in circulating virus and thus hospital burdens, and differences in hospital capacity and standards of care [92–94]. Fig. 1D shows, however, that inter-country variation is more complex and does not appear to be directly related to the

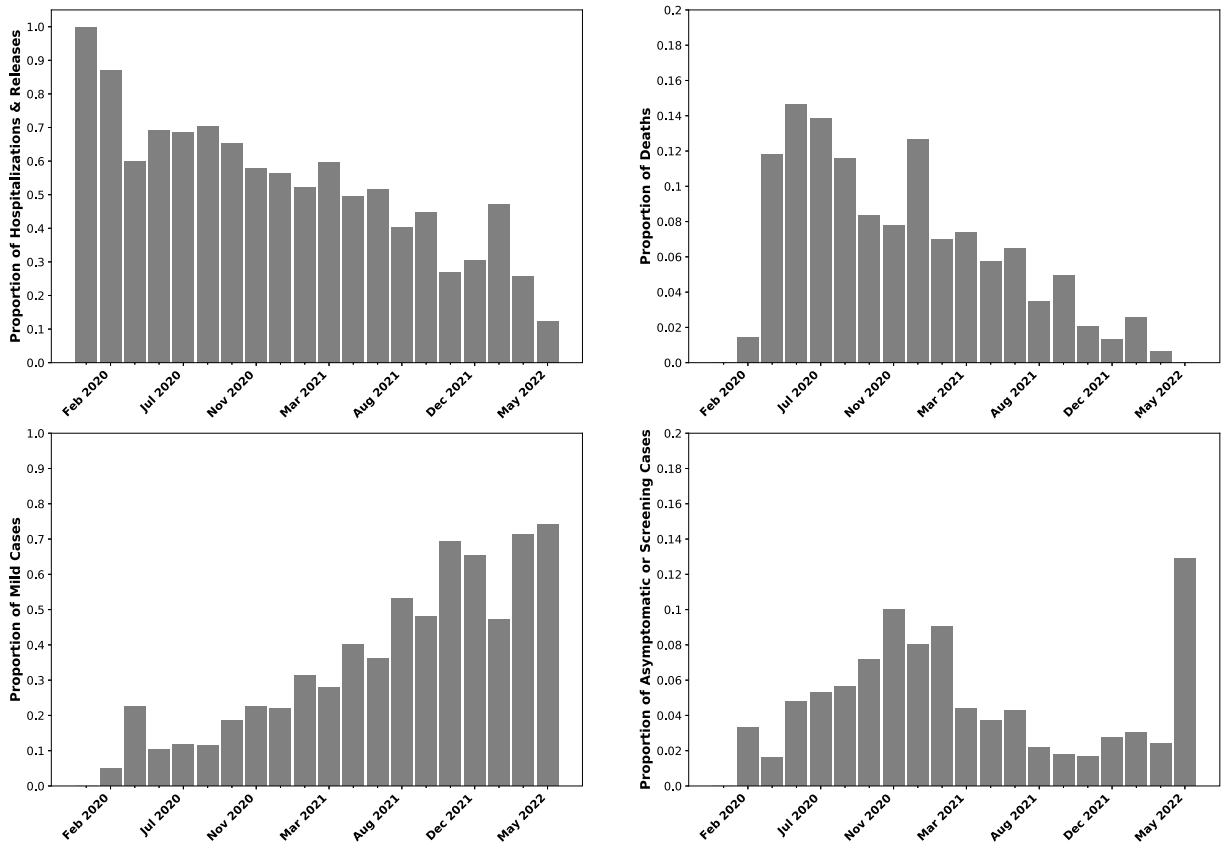

Fig. 3. GISAID patient status metadata trends over time. (A – Upper Left) Fraction of cases categorized as Hospitalized or Released (from hospital) over time, binning dates as indicated by the bars. The definitions of hospitalizations and releases based on patient metadata are provided in 1 and 1. **(B – Upper Right)** Fraction of samples annotated as being from dead individuals in the GISAID patient status metadata field, binned by date as in Panel A. The cases in Panels A and B are collectively classified as Severe. **C - Lower Left** Fraction of cases categorized as Mild according to 1. **(D – Lower Right)** Fraction of cases categorized as Asymptomatic or Screening according to 1. Panels C and D are collectively classified as Mild. The subgroups of Mild and Severe classifications show similar trends, showing that the overall trends in Fig. 1 are not due to changes in how metadata are described and characterized.

aforementioned factors. Some of the data show consistent trends or levels across all time points. For example, essential all records submitted from Hong Kong, and the overwhelming majority of records from Brazil, are classified as Severe (hospitalizations or deaths). Cases from France, which as discussed above is the largest source of sequences, show a decline over time, although with a lot of fluctuation. Cases from neighboring Belgium, by contrast, have been consistently increasing in severity since Summer 2021. Samples from the United States have increased and decreased in severity with no discernible pattern. In sum, samples originating from different countries follow distinct patterns, which means that including country as a feature will likely improve classification. But the country feature will reflect sampling differences over time between countries—whether the sequences are being submitted mostly from hospital settings, or if that is the case at different points in time. While there is variance between samples from a country that is independent from that of other countries, it is at least in large part due to factors that are not relevant to disease outcome.

Accordingly, we hypothesize that sampling variations can be modeled as a random effect in a linear mixed effects model [72,73,95], where country of origin is a random effect group rather than a feature. To test this hypothesis, in the following section we compare mixed effect machine learning to other classification methods for predicting disease severity. The comparison is based on GISAID data from a timeframe when the overall decrease in severity will not confound a model. Otherwise, as discussed above, any model will inevitably make predictions that are not clinically relevant going forward. For example, mutations that emerged in Delta will be found as leading to less severe disease, and thus when they are found in Omicron sublineages, they will be predicted as reducing severity—due to the potential artifact in samples collected during the period when Delta and Alpha coincided discussed above. We therefore limit our analysis of modeling methods to samples collected beginning in July 2021, when the declining trend in cases has stabilized (see Fig. 1B and C). The result is a training data set made up of mostly Delta subvariants, along with a substantial number of Omicron sequences, and smaller numbers of other lineages such as the Gamma (P.1) variant of concern, which has shown some ability to evade neutralizing antibodies [96,97] and may result in increased disease severity [98]. Doing so helps avoid confounders in analyzing the impact of different mutations and combinations of mutations in Delta and Omicron subvariants, allowing for predictions about whether mutations observed in Delta will result in more severe disease if they occur in future Omicron variants.

3.2. Comparison of machine learning methods

To evaluate our hypothesis that a mixed effect modeling approach can be useful for GISAID patient data, where country should be treated as a random effect group, we evaluate the GPBoost mixed effect machine learning method [50] alongside two popular highly efficient ensemble decision tree methods which employ gradient-boosting, XGBoost and LightGBM [53], random forests [51], a well-established ensemble decision tree method, and conventional logistic regression with elastic net regularization [65]. To compare with GPBoost, two feature sets of models are evaluated: (i) using sequence, age, and gender as features, and (ii) using country metadata as an additional feature.

As discussed above, the following analysis focuses on GISAID data starting from July 2021. Models are trained on the samples collected between July 17, 2021 through December 25, 2021 (68,815 in total).

Testing is performed on samples collected entirely after the training period, from December 26, 2021 through the latest-collected sequences from April 10, 2022 (42,420 samples). Although cross-validation is a typical way of evaluating classifiers to avoid overfitting [99,100], in this data set, as shown in Fig. 1 and discussed above, there will be potential clusters of confounding variables at different times. For example, a narrow range of sequence collection dates may correspond to a study of patients with common characteristics, e.g., patients who are all hospitalized, or mildly symptomatic patients from a screening study. Cross-validation will sample time points evenly, and, as such, a classifier may overfit to patterns within the confounders and then appear to perform better than it otherwise would be on a realistic prediction task. As an alternative, we evaluate the classifiers on temporally split data: i.e., we seek to predict disease severity for sequences from a model trained on samples collected earlier. Previous work has confirmed that temporally split test and training sets provide a more realistic evaluation of classification methods, in that methods perform less well when evaluated on a temporally split validation data set than they do using cross-validation [45]. Notably, there is some class imbalance, which is similar between the training and test data sets: 39.2% of the test samples were severe and 37.4% of training were Severe. Class or sample balancing did not substantially affect the results for the methods which allow it, i.e., not GPBoost (data not shown).

Fig. 4 compares machine learning methods by showing aggregate and class-specific test data classification metrics for models trained using the different methods under evaluation. The aggregate metrics shown in Fig. 4 are the accuracy, which is measured as the number of correct class predictions divided by the number of total predictions, and the balanced (weighted average) *F1*-score, which reflects the sensitivity and specificity of the predictions, accounting for the aforementioned class imbalance. The balanced *F1*-score is the harmonic mean of precision (true positives divided by all positive predictions) and recall (true positive rate, i.e., sensitivity). Fig. 4 also shows the class-specific precision and recall, which is a useful measure as to whether methods might underperform on predicting a particular class, such as the minority (Severe) or majority (Mild) class.

Although the performance of the methods varies depending on metric, two trends are clear. First, the best-performing methods are consistently (1) the high-performance gradient boosting decision tree methods, XGBoost, LightGBM, and GPBoost, with class prediction accuracies above 75%. Second, the best-performing methods account for country—either as an independent feature, in the case of XGBoost and LightGBM, or as group level random effects for the mixed effects model trained by GPBoost. Notably, these three methods, unlike classical regression methods, can handle missing information for sequence positions, which are allowed to increase the number of data for training. Fig. 5 further shows the receiver operator characteristic (ROC) curves for the best-performing methods, and reports the area under the curve (AUC), which provides a metric for comparing model performance. Using the AUC metric, XGBoost with country as an independent feature has the highest AUC. It is important to keep in mind, however, that as shown in Fig. 1D, a model that includes country as a feature may be overfitting to consistent sample collection biases. Notably, GPBoost outperforms LightGBM and XGBoost when country is not an independent feature of the latter two models. GPBoost also has a higher AUC than LightGBM and only marginally lower than that of XGBoost.

To further compare the performance of the best-performing models, they can be tested on their ability to predict whether specific sequence mutations affect the relative risk of severe disease. This analysis focuses on two specific spike protein site mutations for which there is substantial evidence from both epidemiological and laboratory studies for increased disease severity: a leucine-to-arginine mutation at position 452 (L452R) and a proline to arginine mutation at position 681 (P681R). These mutations are characteristic of the Delta variant [101]. SARS-CoV-2 with P681R has been found to have higher spike protein

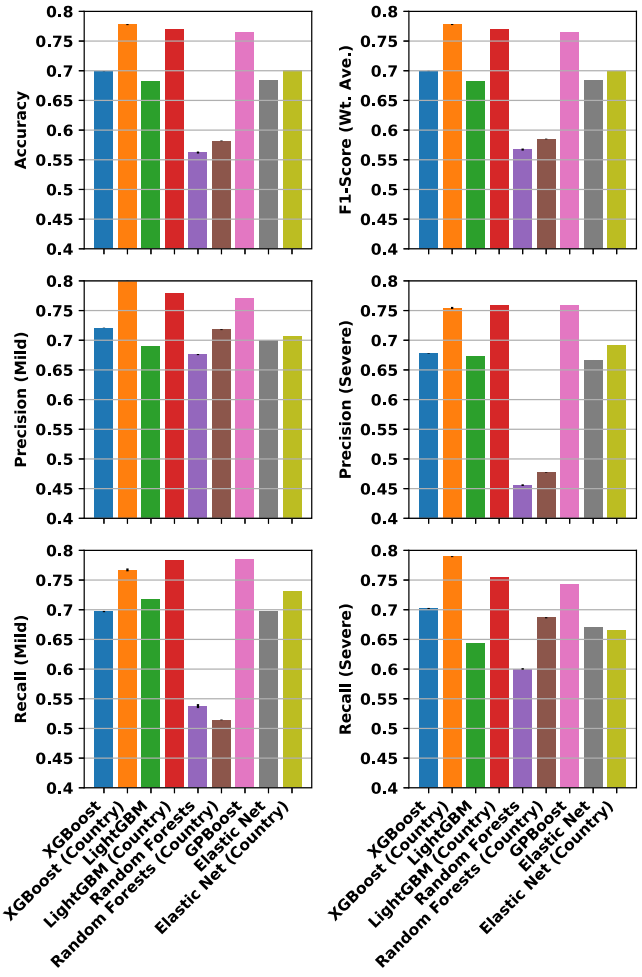

Fig. 4. Comparison of classification metrics for different machine learning methods. Metrics are computed on test samples collected from December 26, 2021 through April 10, 2022, for models trained on samples from July 17 through December 25, 2021. The top row shows, at left, the accuracy of the Mild/Severe classification and, at right, the balanced *F1*-score, which is the harmonic mean of precision (true positives divided by all positive predictions) and recall (true positive rate, i.e., sensitivity). The middle row shows the precision for the Mild and Severe class predictions separately, and the bottom row shows the recall. Metrics are shown for models trained with country metadata used as a feature and without, as indicated in the labeled axes below, except for GPBoost, which takes into account the country metadata by using it as the groups of random effects. All models otherwise use age, gender, and each sequence position as a feature. Error bars show the standard deviations across three runs with different random number seeds, and in some cases are not visible. Statistics for GPBoost are computed based on the mean of the response. GPBoost and LightGBM/XGBoost including country as a feature consistently outperform other methods.

cleavage and viral fusogenicity in vitro, and result in higher pathogenicity in a Syrian hamster animal model [102]. Another study introducing P681R on an Omicron background showed an increase in fusogenicity and synctitia formation, which have been correlated to pathogenicity [103]. The L452R mutation has been found to also increase viral fusogenicity in vitro, and to result in increased infectivity in a mouse lung cell model [104]. Another in vitro study has also shown that L452R resulted in increased spike protein stability, viral fusogenicity and infectivity, and, in turn, increased viral replication [105]. And, the Delta variant, which is characterized by the P681R and L452R mutations, was to result in an increased risk of hospitalizations in epidemiological studies in Denmark [13], England [16], and Canada [14]. Accordingly, a model would be expected to show that a L452R or P681R mutation will result in greater severity.

As an additional validation study, therefore, machine learning models are evaluated on whether they are more likely to a predict a Severe

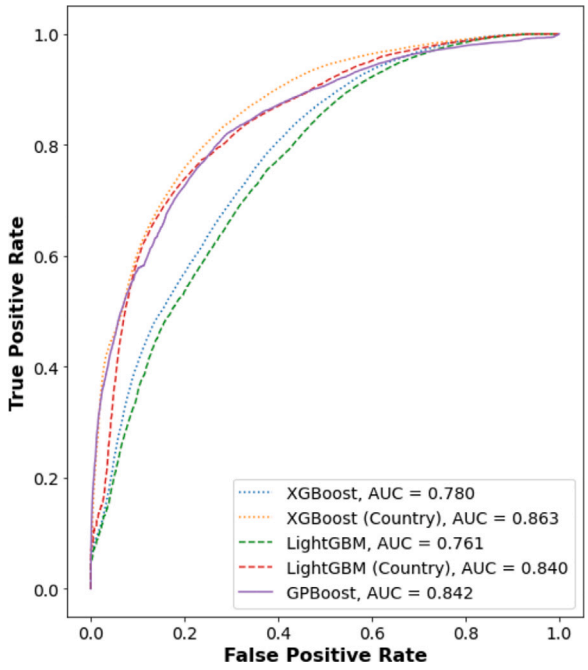

Fig. 5. Receiver operator characteristic curves for best-performing modeling methods. ROC curves were obtained using the `scikit-learn` package version 1.0.2 [62] for test samples and trained models as described for Fig. 4 for XGBoost and LightGBM (with and without country metadata) and GPBoost (using country as a random effects group). The data are shown for one run; run-to-run variation was found to be insignificant. GPBoost performs better than either LightGBM or XGBoost, unless country metadata are used for the latter methods.

classification in the presence of L452R or P681R sequence changes. However, the methods being compared here are decision tree-based methods, which unlike classical logistic regression do not generate coefficients that can be used to analyze individual features. The impact of specific feature changes may be estimated instead. In particular, SHAP values can be utilized in conjunction to provide an estimate of the log-odds for a Severe case given a particular feature value [54]. SHAP values are typically generated for a subset of samples, as it is a computationally intensive process. Fig. 6 shows SHAP dependency plots for samples collected from March 8 through April 10, 2022 (5918 samples). The points in the plots represents the estimated SHAP value (log-odds for a Severe case) for each sample; the color indicates the age of the patient for the sample. This means that SHAP dependency plots show how a specific feature interacts with another feature: the age of the patient in Fig. 6. (As indicated in Fig. 1, age has a significant correlation with disease severity in GISAID patient data, as well as in real-world epidemiological studies.) The SHAP dependency plots represent the potential sequence features at spike protein positions 452 and 681: L (ancestral), leucine, M, methionine, and R, arginine for position 452, and P (ancestral), proline, H, histidine, R, and Y, tyrosine for position 681. (P681H is a common mutation founds in Omicron sequences [106]). The '*' character indicates that there was a missing amino acid at that position in the sampled sequence, likely due to sequencer error, which is treated as missing data by the respective methods. As Fig. 6 shows, GPBoost is the only method which shows an increased SHAP value, or estimated log-odds of a Severe outcome, for the L452R and P681R mutations.

3.3. Predicting the potential severity of emerging omicron variants

A key objective for training a sequence-phenotype model is to be able to predict how novel combinations of mutations – such as the reemergence of a mutation found in a separate lineage – could affect pathogenicity and clinical outcomes. Here, the potential utility of a spike protein sequence-clinical severity prediction model trained on GISAID data is demonstrated for Omicron lineages emerging as significant threats as of May 2022: BA.4 and BA.5, which had become the predominant variants in South Africa and found to be rapidly growing in Portugal [107], and BA.2.12.1, which had accounted for substantial case growth in the United States [108].

The predicted relative severity resulting from different spike sequences may be compared by looking at the relative raw (unrounded) prediction of the model. In the context of this paper, the probability on the logistic curve fit by the model that the binary classification will be 0 or 1. In practice, the class prediction is provided by rounding the model output to 0 (Mild) or 1 (Severe), i.e., to generate the classification metrics shown in Fig. 4. However, as explained above, GISAID data do not provide a realistic measurement of the actual observed probability of severe outcomes, as there are far more hospitalized and deceased patients than real-world hospitalization and CFR data indicate. The quantitative model predictions should be interpreted, therefore, in a relative manner. Accordingly, the raw model output can help in providing relative predictions, but should not be interpreted as an absolute probability of severe disease. In sum, predictions for the aforementioned emerging sublineages may be compared against the predictions for the original Omicron sublineages, BA.1 and BA.2.

Fig. 7 shows the output of the trained GPBoost, LightGBM, and XGBoost models, where the latter two include country as a feature, as shown above in Fig. 6. The sequences used to generate the predictions in Fig. 7 are the most common of those variants found in the GISAID patient data set used in this paper (collected before April 15, 2022), with GISAID accession numbers as provided in the figure caption. An additional BA.2.12.1 sequence collected after the data set used in this paper was separately retrieved from GISAID (accession number EPI_ISL_12048110). As Fig. 7 illustrates, Country has a substantial impact on the predictions made using LightGBM and XGBoost. This sharply limits the utility of LightGBM and XGBoost models as predictive tools.

While it is possible to standardize the country and view the prediction relatively, as shown in 7, the relative difference between variants differs greatly between countries. For a simulated patient sample from the United States, the variants have nearly identical (and very high) predictions, while the predictions for simulated samples from France vary differently, with much more dynamic range. Samples from Mexico are in between. GPBoost models, by contrast, do not vary between countries. The mixed effects model trained by GPBoost does not account for country in grouping only *random* effects. By considering only the mean model response, random effects cancel each other out, and there is only one prediction for any country. Given that, as shown in Fig. 1D, the differences between countries are apparently unrelated to actual local conditions, such as access to treatment, a country-neutral prediction provides a more realistic, and likely more relevant, of the relative increase in severe disease risk associated with a new SARS-CoV-2 variant.

Accounting for the variation between countries, Fig. 7 generally shows that BA.2, BA.2.12.1, BA.4, and BA.5 all have higher predicted severity than BA.1. Notably, a study of infectivity in mouse and hamster models suggested that there is no difference in infectivity, replication, and pathogenicity between BA.1 and BA.2 virus [109]. Another study, however, found greater fusogenicity and replication in nasal epithelial cells studied in vitro, as well as more pathogenicity in a hamster model for BA.2 as compared to BA.1 [110]. Moreover, a recently published population study in England reports that individuals infected with BA.2 reported more symptoms than those with BA.1 [111]. Another study of patients in Italy also reported more symptomatic disease when infected with BA.2 rather than BA.1 [112].

The SHAP method used for feature analysis above can be used to examine in detail how specific features influence the prediction for emerging variants as well [54]. Fig. 8 shows exemplary SHAP

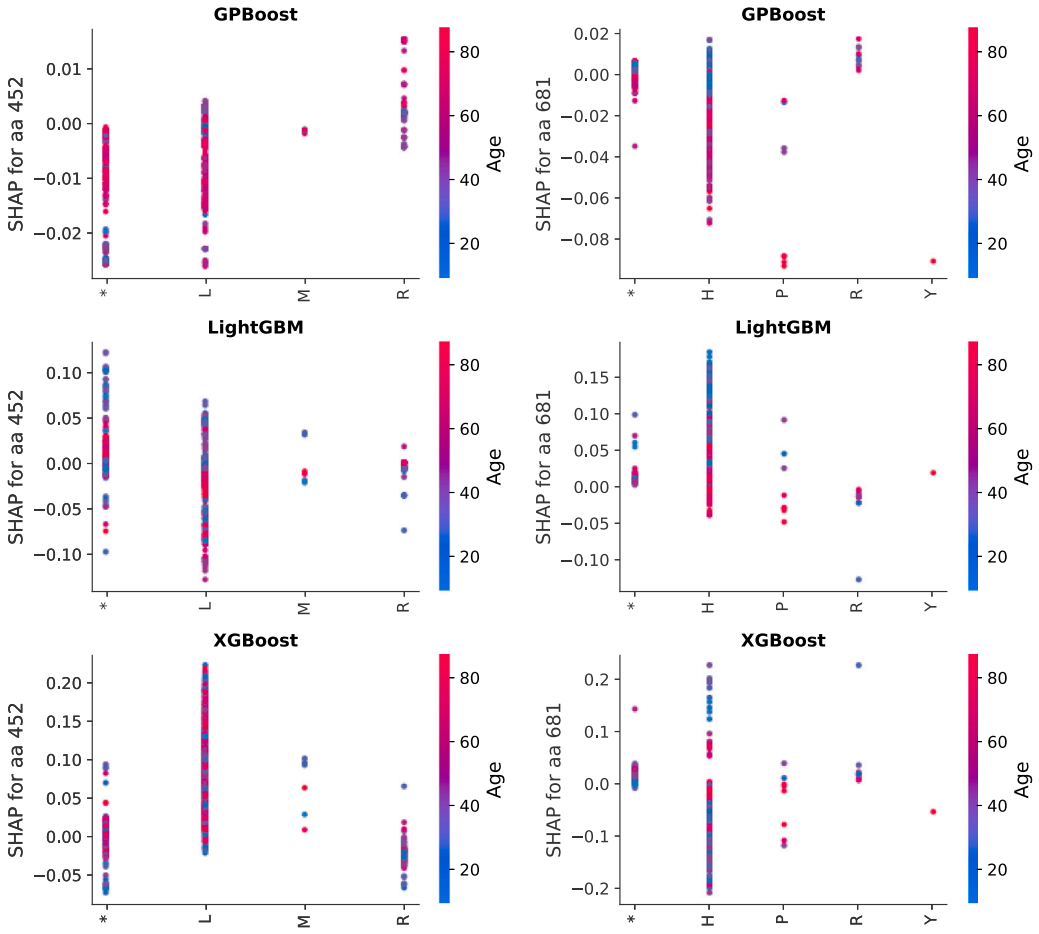

Fig. 6. Comparison of SHAP dependence plots to severity for sequence positions 452 and 681 for the best-performing models. LightGBM and XGBoost with country as a feature are compared to the GPBoost mixed effect model, trained aon data as described in 4. The predicted SHAP values for each of the samples used to generate the SHAP estimate (sequences collected from March 8 through April 10, 2022) are plotted for the 452 and 281 sequence positions in the left and right columns respectively, showing the SHAP values for predictions with sequences of the indicated amino acid at that position, i.e. L (ancestral), leucine, M, methionine, and R, arginine for residue 452; P (ancestral), H, R, and Y for residue 681; and '*' for missing amino acid in the sample. A positive SHAP value indicates that an amino acid change is positively related to increased severity. The interaction of the patient age feature is shown by the coloring of the points, where more red points are from older patients and blue points from younger. GPBoost indicates increased severity as expected from validated experiments of L → R for this time period.

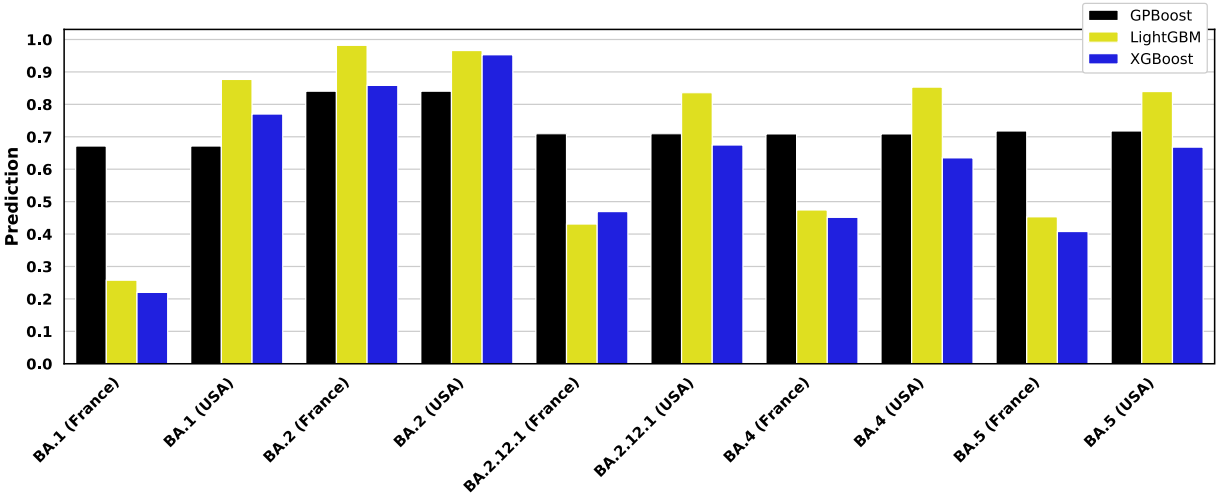

Fig. 7. Predictions of Omicron subvariant severity. Trained GPBoost, LightGBM, and XGBoost models are simulated for representative BA.1, BA.2, BA.2.12.1, BA.4, and BA.5 sequences from a 60 year-old male patient obtained in the United States, France, and Mexico. The GISAID accession numbers of the sequences for the sequences are: EPI_ISL_6590782 (BA.1), EPI_ISL_7852877 (BA.2), EPI_ISL_12048110 (BA.2.12.1), EPI_ISL_11674447 (BA.4), and EPI_ISL_12029894 (BA.5). The predictions shown here are for models trained on training data as shown in Figs. 4 and 6 where country is a feature for LightGBM and XGBoost. The GPBoost predictions shown here are for the mean of the model response, and it does not vary by country, since country is not a fixed effect in the mixed effects model trained using GPBoost. By contrast, LightGBM and XGBoost predictions fluctuate significantly by simulated country. Emerging Omicron subvariants are uniformly predicted to be more severe than BA.1.

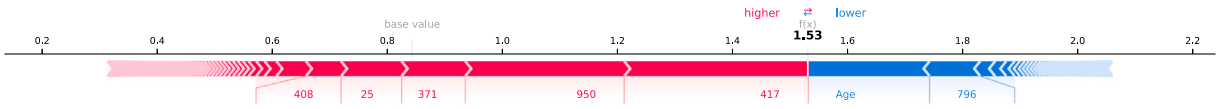

Fig. 8. SHAP force plot showing impact of features on BA.2.12.1 severity prediction by GPBoost. The "force plot" is a visualization which shows, based on SHAP values estimating the log-odds contribution of features to the model prediction, how much a specific feature tends to weigh the decision between binary classes. This plot is based on a simulated 30 year old male patient, and thus the Age feature tends to weigh the model towards a Mild prediction for this sample. Other features tend to weigh towards a more Severe prediction, such as mutations at sites characteristic of BA.2, including positions 371 and 408.

visualization for the GPBoost prediction of the representative BA.2.12.1 sequences shown in Fig. 7 (GISAID accession EPI_ISL_12048110), simulated for a 30 year old male patient. The plot shows how key features tend to make a prediction of greater severity (indicated by an increasing value) or lower severity (decreasing value). In the case of this younger patient, for example, the Age feature tends to reduce the predicted severity. Notably, Fig. 8 suggests that three mutations characteristic of BA.2 influence an increase in predicted severity for BA.2.12.1: a deletion at positions 24 through 26, S371F (serine to phenylalanine), and R408S (arginine to serine) [113]. S371F is a mutation in the receptor binding domain (RBD) of the spike protein which has been shown to be evasive to antibodies [114,115]. While an antibody evasive mutant might not necessarily confer greater severity on an immunonaive patient, given the high rates of vaccination and/or prior infection now, a model based on contemporary GISAID data can be expected to shown greater severity for immune escape variants. While the impact is smaller than for those features shown in Fig. 8, analysis of SHAP values shows that another immune escape change found in BA2.12.1, E484 A, also tends to elevate the severity prediction [116]. Similarly, BA.4 and BA.5 have been found to be more immunoevasive in BA.1, which may also result in increased severe disease among populations with acquired immunity [117,118].

4. Discussion

Global genome repositories like GISAID have the potential to be an unparalleled resource for understanding and quantitatively modeling genotype-phenotype relationships. As the foremost repository for SARS-CoV-2 genome sequences, GISAID offers the largest possible potential data set with the greatest global reach. As a result, GISAID can solve one of the key challenges with biomedical modeling problems: small data set sizes which make them particularly vulnerable to overfitting, because it is often difficult and costly to obtain experimental data [119–121]. The best (and perhaps only real) solution to overfitting is to have more data to develop models. Conventional meta-analyses require searching for relevant studies and parsing through papers with often inconsistent formats and data reporting methods, and they are also limited to published or otherwise documented studies. However, because repositories are generally incorporating multiple studies collected from different sites and under different conditions, heterogeneity is still the key challenge [122]. As Fig. 1 and the accompanying text explain, heterogeneity is a critical problem with GISAID data. The challenges of GISAID source data heterogeneity are particularly exacerbated by the very limited metadata associated with patient samples, even for the small subset for which patient status metadata are available at all. Sequence repository data will be more useful as efforts continue to grow to collect and curate important information about the sample and establish minimum information standards [123]. The results in this paper demonstrate that, accounting for the aforementioned caveats, useful information can be obtained by analyzing the GISAID patient data set. There are three key problems with the data set that analytical and modeling methods need to address.

First, it is hard to robustly define mild and severe cases based on patient status metadata. As an initial matter, metadata entries are often inconsistent between different entries or noisy and hard to interpret reliable (see, e.g., Supplementary Table S1). This paper takes a hierarchical approach to defining mild and severe cases, based on established clinical definitions [63], as described in Supplementary Table S2. However, because of confounding variables like vaccination, therapeutic availability, and prior infection, it has become difficult to estimate the "intrinsic" severity of variants [124]. A particular challenging issue concerns whether hospitalizations should be considered as mild or severe cases, especially given how prevalent they are in the GISAID data set (see, e.g., Fig. 3 and accompanying text). While this paper treats them as severe cases, that definition has become increasingly unreliable as the vaccination has become more prevalent. Studies from multiple sites suggests that as vaccination has increased, more hospitalization patients classified have only tested positive on admission but have mild or no symptoms [125–128]. Moreover, the kinds of sequence variation that lead to more severe clinical outcomes may change due to vaccination. As suggested in Fig. 8 and accompanying text, as the overwhelming majority of individuals in many regions have at least some immunity due to vaccination and/or prior infection, immune escape variants may result inn severe disease because they can evade immune responses that would otherwise rapidly clear the virus and prevent infection. However, such variants may not result in more pathogenicity in immunonaive hosts, and thus would not show more severe outcomes earlier in the pandemic.

Second, conditions have changed over time, as shown in the reduction of case severity over time shown in Fig. 1, which consists of reductions in the proportions of both hospitalizations and deaths (see Fig. 3). While trends of decreasing severity are consistent with improved patient outcomes due to vaccination and improved therapeutics [81–86], they may also reflect changes in sequence collection practices, such as obtaining more sequences and performing more studies based on screening the general public outside of hospital settings. These artifacts can have significant impacts on modeling studies. For example, the models derived in this study similarly indicate that the E1258D spike protein mutation has a significant impact on increasing severity. In addition to our group's previous work, another independent investigation of GISAID terminating in Fall 2021 showed E1258D as the strongest sequence feature in determining the severity prediction [44,47]. While E1258D was observed in one publication as an observed result of a missense mutation, that study did not show any effects for that mutation on increased pathogenicity [129]. In fact, E1258D is only found in 1898 of the over 160,000 samples analyzed in this paper. Of those samples, 1849, or 97.4% originated from Mexico, of which 1772 were hospitalized or released from hospital (i.e. considered severe in most studies), and 76 were deceased. Significantly, the metadata were all consistent, including at the level of capitalization, whereas metadata entries generally showed a high degree of heterogeneity. (Supplementary Table S1 shows all unique entries.) Therefore, it is highly likely that E1258D is either sequence artifact, particularly as it is in the cytoplasmic tail of the spike protein and the result of a missense mutation, and thus potentially an unreliable site for interpreting short-read next generation sequencing technologies [130,131]. In sum, caution must be employed in interpreting any features identified as important.

Third, the region from which sequences are collected can have a significant impact on data due to systematic bias. As the E1258D feature demonstrates, large-scale studies in particular regions may interpret sequence data in such a way that can identify a spurious variant if it is

inconsistent with other studies. As Fig. 1D shows, even though it seems logical to ascribe regional differences in clinical outcomes to factors like vaccination, fluctuations at the country level are either virtually constant or otherwise have no consistent pattern. As such, country-level seem more reflective of how samples are collected and metadata are annotated within countries, which motivates the use of mixed effect models as has been previously used for genotype-phenotype modeling where sample batches affect the data [70–73]. The results in this paper demonstrate that a mixed effect machine learning approach in which countries are groups for random effects can be successful in developing a predictive model. The GPBoost method [50] proves to be fast, effective, and robust to missing data, which suggests that it should be more widely utilized in modeling genetic variation. Notably, as Fig. 4 illustrates, using country as a feature does result in much more accurate models. However, these models are overfitting to country-level trends, as evinced by the predictions graphed in Fig. 7, which show dramatic differences in predictions between different countries. As such, while previous studies of GISAID data have shown that including country metadata as a feature in models provides greater explanatory power [42,43,46], any resulting models are likely overfitting as they are here and will have difficulty being generalized to real-world predictions. Accordingly, while region-level features may appear to result in superior models, they risk creating artifacts. For example, as shown in this paper, including country as a feature results in predictions for the impact of L452R and P681R mutations at odds with epidemiological and in vitro evidence (see Fig. 6). The challenges of country-level variation are heightened by substantial regional imbalances in the GISAID patient data set. The entire GISAID database is fundamentally biased towards Europe, North America, and select countries in Asia and elsewhere, with over half the sample originating in either the United Kingdom or United States as of January 2022 [49]. Within the subset of data with patient status metadata, the biases are similarly idiosyncratic; for example, over 40% of the training and test samples shown in this paper were obtained from France.

In addition to the foregoing issues, the work in this paper has further limitations in scope. GISAID patient metadata omit information about comorbidities known to increase the risk of severe clinical outcomes and mortality, such as chronic disease and obesity [132,133]. Studies have shown that host (patient) genetics may also be significant determinants of infection outcomes. a [134–136] Indeed, a recent study showed many genetic correlates of severe COVID-19 that were also correlates for other chronic conditions associated with heightened severity, with a particular focus on immune-mediated conditions [137]. Epigenetic factors may also be significant [138], as well as the host transcriptome [139]. In addition, to make the work shown here more tractable, we focus on the spike protein. However, there is some evidence that a mutation in the nucleocapsid gene may account for some of Delta's increased severity [140]. Finally, the methods described herein rely on training or fitting to existing databases. Entirely novel mutations will not be accounted for and may result in unpredictable outcomes. However, it may be possible to train models on the predicted or in vitro studies of novel mutations that could emerge in the future, such as those identified in deep mutational scanning and other exploration of the mutational landscape [141–144]. In sum, while there are important caveats to utilizing the GISAID data set as a resource for modeling clinical outcomes based on viral genotypes, it provides the most diverse and largest data set possibility. Any other meta-analyses will inherently suffer from the same kinds of data heterogeneity, and will necessarily be more limited as there is data in GISAID beyond that contained in published reports. The relative success of a mixed effect modeling approach suggests that refining the modeling of group level random effects or otherwise incorporate hidden variables are necessary to account for structural issues in the data. Moreover, having established a proof of concept in this study using logistic regression and boosted decision trees, future work can explore the potential application of deep learning methods, which have proven to be highly useful to genetic sequence to function modeling in other contexts [49,145–147].

5. Conclusion

Despite increasingly widespread vaccination and development of new antiviral therapies, COVID-19 continues to represent a significant threat to human health. The virus also continues to be highly unpredictable. Significant genetic variants of SARS-CoV-2 continue to proliferate, and the risk of severe disease in an emerging variant is a particular concern. A critical tool in staying ahead of the virus can be a predictive model for the risks of severe disease based on viral genotype. Potentially predictive genotype-disease severity models depend on a substantial amount of patient data, which exceeds the capability of conventional epidemiological studies and meta-analyses. Patient data within GISAID, the primary global SARS-CoV-2 sequence repository, therefore, represents a key resource for building predictive models. Unfortunately, GISAID patient metadata are limited, both in number and quality; for example, there is no data on comorbidities or vaccination status. Despite these caveats, it has been previously shown that GISAID patient metadata can be used to develop predictive models. However, until this paper, there has not been a rigorous analysis of potential confounders within the data which may prevent such models from being clinically useful.

As shown in this paper, there are temporal trends in sample collection biases which must be accounted for in model training. Moreover, there are significant differences in sample collection biases between countries. Models are more predictive if they take country-of-origin of sequences into account, but such models are likely overfitting to artifacts in how samples are collected in different countries. This study demonstrates that a superior approach to accounting for variation between the country-of-origin of viral sequence and patient data is to employ mixed effects modeling, where country is treated as a random effect group. Mixed effects modeling can be efficiently implemented for the large number of sequence features analyzed in this paper by using the recently developed GPBoost package, which uses gradient boosted decision trees for fixed effects with performance comparable to XGBoost and conventional LightGBM. This study also presents a novel way to validate genotype-disease severity models for COVID-19: interpreting models to determine whether they are able to show that they can predict the effect of known mutations which affect disease severity. This kind of validation further reinforces the potential superiority of mixed effects methods over conventional logistic regression and boosted decision tree methods. Finally, trained GPBoost genotype-severity models are shown to be able to predict severity of emerging SARS-CoV-2 Omicron variants. For example, the GPBoost model presented in this paper predicts that BA.2 and subsequent Omicron variants may pose a greater risk of severe disease than Omicron BA.1, in line with preliminary epidemiological evidence.

CRediT authorship contribution statement

Bahrad A. Sokhansanj: Conceptualization of this study, Data curation, Methodology, Data analysis, Software, Writing – original draft. **Gail L. Rosen:** Conceptualization of this study, Data analysis, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We gratefully acknowledge the following Authors from the Originating laboratories responsible for obtaining the specimens and the Submitting laboratories where genetic sequence data were generated and shared via the GISAID Initiative, on which this research is based. A table of acknowledgments is located at https://epicov.org/epi3/epi_set/EPI_SET_20220606hk (DOI link:https://doi.org/10.55876/gis8.220606hk). GLR received U.S. National Science Foundation (NSF) grants #1919691, #1936791, and #2107108. The funders had no role in study design, deciding to publish, collecting or analyzing data, or preparing the manuscript. Work reported here was run on hardware supported by Drexel's University Research Computing Facility.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.compbiomed.2022.105969.

References

[1] Y. Shu, J. McCauley, GISAID: Global initiative on sharing all influenza data – from vision to reality, Eurosurveillance 22 (13) (2017) 30494.

[2] S. Khare, C. Gurry, L. Freitas, M.B. Schultz, G. Bach, A. Diallo, N. Akite, J. Ho, R.T. Lee, W. Yeo, S. Maurer-Stroh, GISAID Core Curation Team, GISAID's role in pandemic response, China CDC Wkly. 3 (49) (2021) 1049–1051.

[3] A. O'Toole, E. Scher, A. Underwood, B. Jackson, V. Hill, J.T. McCrone, R. Colquhoun, C. Ruis, K. Abu-Dahab, B. Taylor, C. Yeats, L. du Plessis, D. Maloney, N. Medd, S.W. Attwood, D.M. Aanensen, E.C. Holmes, O.G. Pybus, A. Rambaut, Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool, Virus Evol. 7 (2) (2021) veab064.

[4] A. Rambaut, E.C. Holmes, A. O'Toole, V. Hill, J.T. McCrone, C. Ruis, L. du Plessis, O.G. Pybus, A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology, Nat. Microbiol. 5 (11) (2020) 1403–1407.

[5] D.V. Parums, Editorial: revised world health organization (WHO) terminology for variants of concern and variants of interest of SARS-CoV-2, Med. Sci. Monit. : Int. Med. J. Exp. Clin. Res. 27 (2021) e933622–1–e933622–2.

[6] Y. Liu, J. Rocklöv, The reproductive number of the delta variant of SARS-CoV-2 is far higher compared to the ancestral SARS-CoV-2 virus, J. Travel Med. 28 (7) (2021) taab124.

[7] Y. Liu, J. Liu, B.A. Johnson, H. Xia, Z. Ku, C. Schindewolf, S.G. Widen, Z. An, S.C. Weaver, V.D. Menachery, X. Xie, P.-Y. Shi, Delta spike P681R mutation enhances SARS-CoV-2 fitness over Alpha variant, 2021, http://dx.doi.org/10.1101/2021.08.12.456173.

[8] P. Mlcochova, S.A. Kemp, M.S. Dhar, G. Papa, B. Meng, I.A.T.M. Ferreira, R. Datir, D.A. Collier, A. Albecka, S. Singh, R. Pandey, J. Brown, J. Zhou, N. Goonawardane, S. Mishra, C. Whittaker, T. Mellan, R. Marwal, M. Datta, S. Sengupta, K. Ponnusamy, V.S. Radhakrishnan, A. Abdullahi, O. Charles, P. Chattopadhyay, P. Devi, D. Caputo, T. Peacock, C. Wattal, N. Goel, A. Satwik, R. Vaishya, M. Agarwal, A. Mavousian, J.H. Lee, J. Bassi, C. Silacci-Fegni, C. Saliba, D. Pinto, T. Irie, I. Yoshida, W.L. Hamilton, K. Sato, S. Bhatt, S. Flaxman, L.C. James, D. Corti, L. Piccoli, W.S. Barclay, P. Rakshit, A. Agrawal, R.K. Gupta, SARS-CoV-2 B.1.617.2 Delta variant replication and immune evasion, Nature 599 (7883) (2021) 114–119.

[9] R. Challen, E. Brooks-Pollock, J.M. Read, L. Dyson, K. Tsaneva-Atanasova, L. Danon, Risk of mortality in patients infected with SARS-CoV-2 variant of concern 202012/1: Matched cohort study, BMJ (Clin. Res. Ed.) 372 (2021) n579.

[10] N.G. Davies, S. Abbott, R.C. Barnard, C.I. Jarvis, A.J. Kucharski, J.D. Munday, C.A.B. Pearson, T.W. Russell, D.C. Tully, A.D. Washburne, T. Wenseleers, A. Gimma, W. Waites, K.L.M. Wong, K. van Zandvoort, J.D. Silverman, K. Diaz-Ordaz, R. Keogh, R.M. Eggo, S. Funk, M. Jit, K.E. Atkins, W.J. Edmunds, CMMID COVID-19 Working Group, COVID-19 Genomics UK (COG-UK) Consortium, Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England, Science 372 (6538) (2021) eabg3055.

[11] D. Frampton, T. Rampling, A. Cross, H. Bailey, J. Heaney, M. Byott, R. Scott, R. Sconza, J. Price, M. Margaritis, M. Bergstrom, M.J. Spyer, P.B. Miralhes, P. Grant, S. Kirk, C. Valerio, Z. Mangera, T. Prabhahar, J. Moreno-Cuesta, N. Arulkumaran, M. Singer, G.Y. Shin, E. Sanchez, S.M. Paraskevopoulou, D. Pillay, R.A. McKendry, M. Mirfenderesky, C.F. Houlihan, E. Nastouli, Genomic characteristics and clinical effect of the emergent SARS-CoV-2 B.1.1.7 lineage in London, UK: A whole-genome sequencing and hospital-based cohort study, Lancet Infect. Dis. 21 (9) (2021) 1246–1256.

[12] B. Giles, P. Meredith, S. Robson, G. Smith, A. Chauhan, PACIFIC-19 and COG-UK research groups, The SARS-CoV-2 B.1.1.7 variant and increased clinical severity-the jury is out, Lancet Infect. Dis. 21 (9) (2021) 1213–1214.

[13] P. Bager, J. Wohlfahrt, M. Rasmussen, M. Albertsen, T.G. Krause, Hospitalisation associated with SARS-CoV-2 delta variant in Denmark, Lancet Infect. Dis. 21 (10) (2021) 1351.

[14] D.N. Fisman, A.R. Tuite, Evaluation of the relative virulence of novel SARS-CoV-2 variants: A retrospective cohort study in Ontario, Canada, CMAJ 193 (42) (2021) E1619–E1625.

[15] M.I. Paredes, S.M. Lunn, M. Famulare, L.A. Frisbie, I. Painter, R. Burstein, P. Roychoudhury, H. Xie, S.A. Mohamed Bakhash, R. Perez, M. Lukes, S. Ellis, S. Sathees, P.C. Mathias, A. Greninger, L.M. Starita, C.D. Frazar, E. Ryke, W. Zhong, L. Gamboa, M. Threlkeld, J. Lee, D.A. Nickerson, D.L. Bates, M.E. Hartman, E. Haugen, T.N. Nguyen, J.D. Richards, J.L. Rodriguez, J.A. Stamatoyannopoulos, E. Thorland, G. Melly, P.E. Dykema, D.C. MacKellar, H.K. Gray, A. Singh, J.M. Peterson, D. Russell, L.M. Torres, S. Lindquist, T. Bedford, K.J. Allen, H.N. Oltean, Associations between SARS-CoV-2 variants and risk of COVID-19 hospitalization among confirmed cases in Washington State: A retrospective cohort study, 2021, http://dx.doi.org/10.1101/2021.09.29.21264272.

[16] K.A. Twohig, T. Nyberg, A. Zaidi, S. Thelwall, M.A. Sinnathamby, S. Aliabadi, S.R. Seaman, R.J. Harris, R. Hope, J. Lopez-Bernal, E. Gallagher, A. Charlett, D.D. Angelis, A.M. Presanis, G. Dabrera, C. Koshy, A. Ash, E. Wise, N. Moore, M. Mori, N. Cortes, J. Lynch, S. Kidd, D. Fairley, T. Curran, J. McKenna, H. Adams, C. Fraser, T. Golubchik, D. Bonsall, M. Hassan-Ibrahim, C. Malone, B. Cogger, M. Wantoch, N. Reynolds, B. Warne, J. Maksimovic, K. Spellman, K. McCluggage, M. John, R. Beer, S. Afifi, S. Morgan, A. Marchbank, A. Price, C. Kitchen, H. Gulliver, I. Merrick, J. Southgate, M. Guest, R. Munn, T. Workman, T. Connor, W. Fuller, C. Bresner, L. Snell, A. Patel, T. Charalampous, G. Nebbia, R. Batra, J. Edgeworth, S. Robson, A. Beckett, D. Aanensen, A. Underwood, C. Yeats, K. Abudahab, B. Taylor, M. Menegazzo, G. Clark, W. Smith, M. Khakh, V. Fleming, M. Lister, H. Howson-Wells, L. Berry, T. Boswell, A. Joseph, I. Willingham, C. Jones, C. Holmes, P. Bird, T. Helmer, K. Fallon, J. Tang, V. Raviprakash, S. Campbell, N. Sheriff, V. Blakey, L.-A. Williams, M. Loose, N. Holmes, C. Moore, M. Carlile, V. Wright, F. Sang, J. Debebe, F. Coll, A. Signell, G. Betancor, H. Wilson, S. Eldirdiri, A. Kenyon, T. Davis, O. Pybus, L. du Plessis, A. Zarebski, J. Raghwani, M. Kraemer, S. Francois, S. Attwood, T. Vasylyeva, M.E. Zamudio, B. Gutierrez, M.E. Torok, W. Hamilton, I. Goodfellow, G. Hall, A. Jahun, Y. Chaudhry, M. Hosmillo, M. Pinckert, I. Georgana, S. Moses, H. Lowe, L. Bedford, J. Moore, S. Stonehouse, C. Fisher, A. Awan, J. BoYes, J. Breuer, K. Harris, J. Brown, D. Shah, L. Atkinson, J. Lee, N. Storey, F. Flaviani, A. Alcolea-Medina, R. Williams, G. Vernet, M. Chapman, L. Levett, J. Heaney, W. Chatterton, M. Pusok, L. Xu-McCrae, D. Smith, M. Bashton, G. Young, A. Holmes, P. Randell, A. Cox, P. Madona, F. Bolt, J. Price, S. Mookerjee, M. Ragonnet-Cronin, F.F. Nascimento, D. Jorgensen, I. Siveroni, R. Johnson, O. Boyd, L. Geidelberg, E. Volz, A. Rowan, G. Taylor, K. Smollett, N. Loman, J. Quick, C. McMurray, J. Stockton, S. Nicholls, W. Rowe, R. Poplawski, A. McNally, R.M. Nunez, J. Mason, T. Robinson, E. O'Toole, J. Watts, C. Breen, A. Cowell, G. Sluga, N. Machin, S. Ahmad, R. George, F. Halstead, V. Sivaprakasam, W. Hogsden, C. Illingworth, C. Jackson, E. Thomson, J. Shepherd, P. Asamaphan, M. Niebel, K. Li, R. Shah, N. Jesudason, L. Tong, A. Broos, D. Mair, J. Nichols, S. Carmichael, K. Nomikou, E. Aranday-Cortes, N. Johnson, I. Starinskij, A.d.S. Filipe, D. Robertson, R. Orton, J. Hughes, S. Vattipally, J. Singer, S. Nickbakhsh, A. Hale, L. Macfarlane-Smith, K. Harper, H. Carden, Y. Taha, B. Payne, S. Burton-Fanning, S. Waugh, J. Collins, G. Eltringham, S. Rushton, S. O'Brien, A. Bradley, A. Maclean, G. Mollett, R. Blacow, K. Templeton, M. McHugh, R. Dewar, E. Wastenge, S. Dervisevic, R. Stanley, E. Meader, L. Coupland, L. Smith, C. Graham, E. Barton, D. Padgett, G. Scott, E. Swindells, J. Greenaway, A. Nelson, C. McCann, W. Yew, M. Andersson, T. Peto, A. Justice, D. Eyre, D. Crook, T. Sloan, N. Duckworth, S. Walsh, A. Chauhan, S. Glaysher, K. Bicknell, S. Wyllie, S. Elliott, A. Lloyd, R. Impey, N. Levene, L. Monaghan, D. Bradley, T. Wyatt, E. Allara, C. Pearson, H. Osman, A. Bosworth, E. Robinson, P. Muir, I. Vipond, R. Hopes, H. Pymont, S. Hutchings, M. Curran, S. Parmar, A. Lackenby, T. Mbisa, S. Platt, S. Miah, D. Bibby, C. Manso, J. Hubb, M. Chand, G. Dabrera, M. Ramsay, D. Bradshaw, A. Thornton, R. Myers, U. Schaefer, N. Groves, E. Gallagher, D. Lee, D. Williams, N. Ellaby, I. Harrison, H. Hartman, N. Manesis, V. Patel, C. Bishop, V. Chalker, J. Ledesma, K. Twohig, M. Holden, S. Shaaban, A. Birchley, A. Adams, A. Davies, A. Gaskin, A. Plimmer, B. Gatica-Wilcox, C. McKerr, C. Moore, C. Williams, D. Heyburn, E.D. Lacy, E. Hilvers, F. Downing, G. Shankar, H. Jones, H. Asad, J. Coombes, J. Watkins, J. Evans, L. Fina, L. Gifford, L. Gilbert, L. Graham, M. Perry, M. Morgan, M. Bull, M. Cronin, N. Pacchiarini, N. Craine, R. Jones, R. Howe, S. Corden, S. Rey, S. Kumziene-SummerhaYes, S. Taylor, S. Cottrell, S. Jones, S. Edwards, J. O'Grady, A. Page, A. Mather, D. Baker, S. Rudder, A. Aydin, G. Kay, A. Trotter, N.-F. Alikhan, L.d.O. Martins, T. Le-Viet, L. Meadows, A. Casey, L. Ratcliffe, D. Simpson, Z. Molnar, T. Thompson, E. Acheson, J. Masoli, B. Knight, S. Ellard, C. Auckland, C. Jones, T. Mahungu, D. Irish-Tavares, T. Haque, J. Hart, E. Witele, M. Fenton, A. Dadrah, A. Symmonds, T. Saluja, Y. Bourgeois, G. Scarlett, K. Loveson, S. Goudarzi, C. Fearn, K. Cook, H. Dent, H. Paul, D. Partridge, M. Raza, C. Evans, K. Johnson, S. Liggett, P. Baker, S. Bonner, S. Essex, R. Lyons, K. Saeed, A. Mahanama, B. Samaraweera, S. Silveira, E. Pelosi, E. Wilson-Davies, R. Williams, M. Kristiansen, S. Roy, C. Williams, M. Cotic, N. Bayzid, A. Westhorpe, J. Hartley, R. Jannoo, H. Lowe, A. Karamani, L.

Ensell, J. Prieto, S. Jeremiah, D. Grammatopoulos, S. Pandey, L. Berry, K. Jones, A. Richter, A. Beggs, A. Best, B. Percival, J. Mirza, O. Megram, M. Mayhew, L. Crawford, F. Ashcroft, E. Moles-Garcia, N. Cumley, C. Smith, G. Bucca, A. Hesketh, B. Blane, S. Girgis, D. Leek, S. Sridhar, S. Forrest, C. Cormie, H. Gill, J. Dias, E. Higginson, M. Maes, J. Young, L. Kermack, R. Gupta, C. Ludden, S. Peacock, S. Palmer, C. Churcher, N. Hadjirin, A. Carabelli, E. Brooks, K. Smith, K. Galai, G. McManus, C. Ruis, R. Davidson, A. Rambaut, T. Williams, C. Balcazar, M. Gallagher, A. O'Toole, S. Rooke, V. Hill, K. Williamson, T. Stanton, S. Michell, C. Bewshea, B. Temperton, M. Michelsen, J. Warwick-Dugdale, R. Manley, A. Farbos, J. Harrison, C. Sambles, D. Studholme, A. Jeffries, L. Jackson, A. Darby, J. Hiscox, S. Paterson, M. Iturriza-Gomara, K. Jackson, A. Lucaci, E. Vamos, M. Hughes, L. Rainbow, R. Eccles, C. Nelson, M. Whitehead, L. Turtle, S. Haldenby, R. Gregory, M. Gemmell, C. Wierzbicki, H. Webster, T. de Silva, N. Smith, A. Angyal, B. Lindsey, D. Groves, L. Green, D. Wang, T. Freeman, M. Parker, A. Keeley, P. Parsons, R. Tucker, R. Brown, M. Wyles, M. Whiteley, P. Zhang, M. Gallis, S. Louka, C. Constantinidou, M. Unnikrishnan, S. Ott, J. Cheng, H. Bridgewater, L. Frost, G. Taylor-Joyce, R. Stark, L. Baxter, M. Alam, P. Brown, D. Aggarwal, A. Cerda, T. Merrill, R. Wilson, P. McClure, J. Chappell, T. Tsoleridis, J. Ball, D. Buck, J. Todd, A. Green, A. Trebes, G. MacIntyre-Cockett, M. de Cesare, A. Alderton, R. Amato, C. Ariani, M. Beale, C. Beaver, K. Bellis, E. Betteridge, J. Bonfield, J. Danesh, M. Dorman, E. Drury, B. Farr, L. Foulser, S. Goncalves, S. Goodwin, M. Gourtovaia, E. Harrison, D. Jackson, D. Jamrozy, I. Johnston, L. Kane, S. Kay, J.-P. Keatley, D. Kwiatkowski, C. Langford, M. Lawniczak, L. Letchford, R. Livett, S. Lo, I. Martincorena, S. McGuigan, R. Nelson, S. Palmer, N. Park, M. Patel, L. Prestwood, C. Puethe, M. Quail, S. Rajatileka, C. Scott, L. Shirley, J. Sillitoe, M.S. Chapman, S. Thurston, G. Tonkin-Hill, D. Weldon, D. Rajan, I. Bronner, L. Aigrain, N. Redshaw, S. Lensing, R. Davies, A. Whitwham, J. Liddle, K. Lewis, J. Tovar-Corona, S. Leonard, J. Durham, A. Bassett, S. McCarthy, R. Moll, K. James, K. Oliver, A. Makunin, J. Barrett, R. Gunson, Hospital admission and emergency care attendance risk for SARS-CoV-2 delta (B.1.617.2) compared with alpha (B.1.1.7) variants of concern: A cohort study, Lancet Infect. Dis. 22 (1) (2022) 35–42.

[17] M.-A. Davies, R. Kassanjee, P. Rosseau, E. Morden, L. Johnson, W. Solomon, N.-Y. Hsiao, H. Hussey, G. Meintjes, M. Paleker, T. Jacobs, P. Raubenheimer, A. Heekes, P. Dane, J.-L. Bam, M. Smith, W. Preiser, D. Pienaar, M. Mendelson, J. Naude, N. Schrueder, A. Mnguni, S.L. Roux, K. Murie, H. Prozesky, H. Mahomed, L. Rossouw, S. Wasserman, D. Maughan, L. Boloko, B. Smith, J. Taljaard, G. Symons, N. Ntusi, A. Parker, N. Wolter, W. Jassat, C. Cohen, R. Lessells, R.J. Wilkinson, J. Arendse, S. Kariem, M. Moodley, K. Vallabhjee, M. Wolmarans, K. Cloete, A. Boulle, Africa, On behalf of the Western Cape and South African National Departments of Health in collaboration with the National Institute for Communicable Diseases in South, Outcomes of laboratory-confirmed SARS-CoV-2 infection in the Omicron-driven fourth wave compared with previous waves in the Western Cape Province, South Africa, 2022, http://dx.doi.org/10.1101/2022.01.12.22269148.

[18] P. Bager, J. Wohlfahrt, S. Bhatt, M. Stegger, R. Legarth, C.H. Møller, R.L. Skov, P. Valentiner-Branth, M. Voldstedlund, T.K. Fischer, L. Simonsen, N.S. Kirkby, M.K. Thomsen, K. Spiess, E. Marving, N.B. Larsen, T. Lillebaek, H. Ullum, K. Mølbak, T.G. Krause, S.M. Edslev, R.N. Sieber, A.C. Ingham, M. Overvad, M.A. Gram, F.K. Lomholt, L. Hallundbæk, C.H. Espensen, S. Gubbels, M. Karakis, K.L. Møller, S.S. Olsen, Z.B. Harboe, C.K. Johannesen, M. van Wijhe, J.G. Holler, R.B.C. Dessau, M.B. Friis, D. Fuglsang-Damgaard, M. Pinholt, T.V. Sydenham, J.E. Coia, E.S. Marmolin, A. Fomsgaard, J. Fonager, M. Rasmussen, A. Cohen, Risk of hospitalisation associated with infection with SARS-CoV-2 omicron variant versus delta variant in Denmark: An observational cohort study, Lancet Infect. Dis. (2022).

[19] L. Wang, N.A. Berger, P.B. Davis, D.C. Kaelber, N.D. Volkow, R. Xu, Comparison of outcomes from COVID infection in pediatric and adult patients before and after the emergence of Omicron, 2022, http://dx.doi.org/10.1101/2021.12.30.21268495.

[20] J.A. Lewnard, V.X. Hong, M.M. Patel, R. Kahn, M. Lipsitch, S.Y. Tartof, Clinical outcomes among patients infected with Omicron (B.1.1.529) SARS-CoV-2 variant in Southern California, 2022, http://dx.doi.org/10.1101/2022.01.11.22269045.

[21] N. Ferguson, A. Ghani, W. Hinsley, E. Volz, Report 50 - Hospitalisation risk for Omicron cases in England, 2021, http://www.imperial.ac.uk/medicine/departments/school-public-health/infectious-disease-epidemiology/mrc-global-infectious-disease-analysis/covid-19/report-50-severity-omicron/.

[22] T. Nyberg, N.M. Ferguson, S.G. Nash, H.H. Webster, S. Flaxman, N. Andrews, W. Hinsley, J.L. Bernal, M. Kall, S. Bhatt, P. Blomquist, A. Zaidi, E. Volz, N.A. Aziz, K. Harman, S. Funk, S. Abbott, T. Nyberg, N.M. Ferguson, S.G. Nash, H.H. Webster, S. Flaxman, N. Andrews, W. Hinsley, J.L. Bernal, M. Kall, S. Bhatt, P. Blomquist, A. Zaidi, E. Volz, N.A. Aziz, K. Harman, S. Funk, S. Abbott, R. Hope, A. Charlett, M. Chand, A.C. Ghani, S.R. Seaman, G. Dabrera, D.D. Angelis, A.M. Presanis, S. Thelwall, R. Hope, A. Charlett, M. Chand, A.C. Ghani, S.R. Seaman, G. Dabrera, D.D. Angelis, A.M. Presanis, S. Thelwall, Comparative analysis of the risks of hospitalisation and death associated with SARS-CoV-2 Omicron (B.1.1.529) and Delta (B.1.617.2) variants in England: A cohort study, Lancet 399 (10332) (2022) 1303–1312.

[23] B. Meng, I. Ferreira, A. Abdullahi, S.A. Kemp, N. Goonawardane, G. Papa, S. Fatihi, O. Charles, D. Collier, J. Choi, J.H. Lee, P. Mlcochova, L. James, R. Doffinger, L. Thukral, K. Sato, R.K. Gupta, CITIID-NIHR BioResource COVID-19 Collaboration, The Genotype to Phenotype Japan (G2P-Japan) Consortium, SARS-CoV-2 Omicron spike mediated immune escape, infectivity and cell-cell fusion, 2021, http://dx.doi.org/10.1101/2021.12.17.473248.

[24] H. Zhao, L. Lu, Z. Peng, L.-L. Chen, X. Meng, C. Zhang, J.D. Ip, W.-M. Chan, A.W.-H. Chu, K.-H. Chan, D.-Y. Jin, H. Chen, K.-Y. Yuen, K.K.-W. To, SARS-CoV-2 Omicron variant shows less efficient replication and fusion activity when compared with Delta variant in TMPRSS2-expressed cells, Emerg. Microb. Infect. 11 (1) (2022) 277–283.

[25] R. Abdelnabi, C.S.-Y. Foo, X. Zhang, V. Lemmens, P. Maes, B. Slechten, J. Raymenants, E. Andre, B. Weynand, K. Dallmeier, J. Neyts, The Omicron (B.1.1.529) SARS-CoV-2 variant of concern does not readily infect Syrian hamsters, 2021, http://dx.doi.org/10.1101/2021.12.24.474086.

[26] K.A. Ryan, R.J. Watson, K.R. Bewley, C.A. Burton, O. Carnell, B.E. Cavell, A.R. Challis, N.S. Coombes, K. Emery, R. Fell, S.A. Fotheringham, K.E. Gooch, K. Gowan, A. Handley, D.J. Harris, R. Humphreys, R. Johnson, D. Knott, S. Lister, D. Morley, D. Ngabo, K.L. Osman, J. Paterson, E.J. Penn, S.T. Pullen, K.S. Richards, I. Shaik, S. Summers, S.R. Thomas, T. Weldon, N.R. Wiblin, R. Vipond, B. Hallis, S.G.P. Funnell, Y. Hall, Convalescence from prototype SARS-CoV-2 protects Syrian hamsters from disease caused by the Omicron variant, 2021, http://dx.doi.org/10.1101/2021.12.24.474081.

[27] D. Planas, D. Veyer, A. Baidaliuk, I. Staropoli, F. Guivel-Benhassine, M.M. Rajah, C. Planchais, F. Porrot, N. Robillard, J. Puech, M. Prot, F. Gallais, P. Gantner, A. Velay, J. Le Guen, N. Kassis-Chikhani, D. Edriss, L. Belec, A. Seve, L. Courtellemont, H. Péré, L. Hocqueloux, S. Fafi-Kremer, T. Prazuck, H. Mouquet, T. Bruel, E. Simon-Lorière, F.A. Rey, O. Schwartz, Reduced sensitivity of SARS-CoV-2 variant Delta to antibody neutralization, Nature 596 (7871) (2021) 276–280.

[28] R.N. Tasakis, G. Samaras, A. Jamison, M. Lee, A. Paulus, G. Whitehouse, L. Verkoczy, F.N. Papavasiliou, M. Diaz, SARS-CoV-2 variant evolution in the United States: High accumulation of viral mutations over time likely through serial Founder Events and mutational bursts, PLOS ONE 16 (7) (2021) e0255169.

[29] A. Baj, F. Novazzi, F. Drago Ferrante, A. Genoni, E. Tettamanzi, G. Catanoso, D. Dalla Gasperina, F. Dentali, D. Focosi, F. Maggi, Spike protein evolution in the SARS-CoV-2 Delta variant of concern: A case series from Northern Lombardy, Emerg. Microb. Infect. 10 (1) (2021) 2010–2015.

[30] A. Baj, F. Novazzi, R. Pasciuta, A. Genoni, F.D. Ferrante, M. Valli, M. Partenope, R. Tripiciano, A. Ciserchia, G. Catanoso, D. Focosi, F. Maggi, Breakthrough infections of E484K-Harboring SARS-CoV-2 Delta Variant, Lombardy, Italy, Emerg. Infect. Diseases 27 (12) (2021) 3180–3182.

[31] L. Chen, M.C. Zody, C. Di Germanio, R. Martinelli, J.R. Mediavilla, M.H. Cunningham, K. Composto, K.F. Chow, M. Kordalewska, A. Corvelo, D.M. Oschwald, S. Fennessey, M. Zetkulic, S. Dar, Y. Kramer, B. Mathema, S. Germer, M. Stone, G. Simmons, M.P. Busch, T. Maniatis, D.S. Perlin, B.N. Kreiswirth, Emergence of multiple SARS-CoV-2 antibody escape variants in an immunocompromised host undergoing convalescent plasma treatment, mSphere 6 (4) (2021) e0048021.

[32] P. Arora, L. Zhang, C. Rocha, A. Sidarovich, A. Kempf, S. Schulz, A. Cossmann, B. Manger, E. Baier, B. Tampe, O. Moerer, S. Dickel, A. Dopfer-Jablonka, H.-M. Jäck, G.M.N. Behrens, M.S. Winkler, S. Pöhlmann, M. Hoffmann, Comparable neutralisation evasion of SARS-CoV-2 Omicron subvariants BA.1, BA.2, and BA.3, Lancet Infect. Dis. (2022) S1473–3099(22)00224–9.

[33] J. Ou, W. Lan, X. Wu, T. Zhao, B. Duan, P. Yang, Y. Ren, L. Quan, W. Zhao, D. Seto, J. Chodosh, Z. Luo, J. Wu, Q. Zhang, Tracking SARS-CoV-2 Omicron diverse spike gene mutations identifies multiple inter-variant recombination events, Signal Transduct. Target. Therapy 7 (1) (2022) 138.

[34] C. Chakraborty, M. Bhattacharya, A.R. Sharma, K. Dhama, Recombinant SARS-CoV-2 variants XD, XE, and XF: The emergence of recombinant variants requires an urgent call for research - Correspondence, Int. J. Surg. (London, England) 102 (2022) 106670.

[35] G.S. Dite, N.M. Murphy, R. Allman, Development and validation of a clinical and genetic model for predicting risk of severe COVID-19, Epidemiol. Infect. 149 (2021) e162.

[36] G.S. Dite, N.M. Murphy, R. Allman, An integrated clinical and genetic model for predicting risk of severe COVID-19: A population-based case-control study, PLoS One 16 (2) (2021) e0247205.

[37] P. Aiewsakun, P. Wongtrakoongate, Y. Thawornwattana, S. Hongeng, A. Thitithanyanont, SARS-CoV-2 genetic variations associated with COVID-19 severity, MedRxiv (2020).

[38] S. SeyedAlinaghi, P. Mirzapour, O. Dadras, Z. Pashaei, A. Karimi, M. MohsseniPour, M. Soleymanzadeh, A. Barzegary, A.M. Afsahi, F. Vahedi, A. Shamsabadi, F. Behnezhad, S. Saeidi, E. Mehraeen, S. Jahanfar, Characterization of SARS-CoV-2 different variants and related morbidity and mortality: A systematic review., Eur. J. Med. Res. 26 (1) (2021) 51.

[39] S.K. Biswas, S.R. Mudi, Spike protein D614G and RdRp P323L: The SARS-CoV-2 mutations associated with severity of COVID-19, Genom. Inform. 18 (4) (2020) e44.

[40] R. Laskar, S. Ali, Differential mutation profile of SARS-CoV-2 proteins across deceased and asymptomatic patients., Chem. Biol. Interact. 347 (2021) 109598.

[41] J. Clauwaert, G. Menschaert, W. Waegeman, E. Dumonteil, D. Fusco, A. Drouin, C. Herrera, F.P. Esper, Y.-W. Cheng, T.M. Adhikari, Z.J. Tu, D. Li, E.A. Li, D.H. Farkas, G.W. Procop, J.S. Ko, T.A. Chan, L. Jehi, B.P. Rubin, J. Li, D.N. Fisman, A.R. Tuite, S.M. Hamed, W.F. Elkhatib, A.S. Khairalla, A.M. Noreddin, R. Sarkar, M. Chawla-Sarkar, S. Majumdar, M. Lo, S. Chattopadhyay, F. Schmidt, Y. Weisblum, M. Rutkowska, D. Poston, J. DaSilva, F. Zhang, E. Bednarski, A. Cho, D.J. Schaefer-Babajew, C. Gaebler, M. Caskey, M.C. Nussenzweig, T. Hatziioannou, P.D. Bieniasz, Geographical and temporal distribution of SARS-CoV-2 globally: An attempt to correlate case fatality rate with the circulating dominant SARS-CoV-2 clades, MedRxiv 193 (42) (2021) 2021.05.25.21257434.

[42] S.M. Hamed, W.F. Elkhatib, A.S. Khairalla, A.M. Noreddin, Global dynamics of SARS-CoV-2 clades and their relation to COVID-19 epidemiology, Sci. Rep. 11 (1) (2021) 8435.

[43] J.D. Voss, M. Skarzynski, E.M. McAuley, E.J. Maier, T. Gibbons, A.C. Fries, R.R. Chapleau, Variants in SARS-CoV-2 associated with mild or severe outcome, Evol. Med. Public Health 9 (1) (2021) 267–275.

[44] R. Agarwal, T. Leblond, E.M. McAuley, E.J. Maier, M. Skarzynski, J.D. Voss, S. Sozhamannan, Linking genotype to phenotype: Further exploration of mutations in SARS-CoV-2 associated with mild or severe outcomes - SARS-CoV-2 coronavirus, 2022, https://virological.org/t/linking-genotype-to-phenotype-further-exploration-of-mutations-in-sars-cov-2-associated-with-mild-or-severe-outcomes/794.

[45] S. Nagpal, N.K. Pinna, D. Srivastava, R. Singh, S.S. Mande, (Machine) learning the mutation signatures of SARS-CoV-2: A primer for predictive prognosis, 2021, http://dx.doi.org/10.1101/2021.08.30.458244.

[46] S. Sawmya, A. Saha, S. Tasnim, M. Toufikuzzaman, N. Anjum, A.H.M. Rafid, M.S. Rahman, M.S. Rahman, Analyzing hCov genome sequences: Predicting virulence and mutation, 2021, http://dx.doi.org/10.1101/2020.06.03.131987.

[47] B.A. Sokhansanj, Z. Zhao, G.L. Rosen, Interpretable and predictive deep modeling of the SARS-CoV-2 spike protein sequence, 2021, http://dx.doi.org/10.1101/2021.12.26.21268414.

[48] F. Obermeyer, S.F. Schaffner, M. Jankowiak, N. Barkas, J.D. Pyle, D.J. Park, B.L. MacInnis, J. Luban, P.C. Sabeti, J.E. Lemieux, Analysis of 2.1 million SARS-CoV-2 genomes identifies mutations associated with transmissibility, medRxiv (2021) 2021.09.07.21263228.

[49] B.A. Sokhansanj, G.L. Rosen, Mapping data to deep understanding: Making the most of the deluge of SARS-CoV-2 genome sequences, mSystems 7 (2) (2022) e00035–22.

[50] F. Sigrist, Gaussian process boosting, 2021, arXiv:2004.02653 [cs, stat].

[51] B.A. Goldstein, E.C. Polley, F.B.S. Briggs, Random forests for genetic association studies, Stat. Appl. Genet. Mol. Biol. 10 (1) (2011) 32.

[52] T. Chen, C. Guestrin, XGBoost: A scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, in: KDD '16, Association for Computing Machinery, New York, NY, USA, 2016, pp. 785–794, http://dx.doi.org/10.1145/2939672.2939785.

[53] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T.-Y. Liu, LightGBM: A highly efficient gradient boosting decision tree, in: Advances in Neural Information Processing Systems, Vol. 30, Curran Associates, Inc., 2017, pp. 3146–3154.

[54] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: Advances in Neural Information Processing Systems, Vol. 30, Curran Associates, Inc., 2017, pp. 4768–4777.

[55] T.S. Pillay, Gene of the month: The 2019-nCoV/SARS-CoV-2 novel coronavirus spike protein, J. Clin. Pathol. 73 (7) (2020) 366.

[56] A.C. Walls, Y.-J. Park, M.A. Tortorici, A. Wall, A.T. McGuire, D. Veesler, Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein, Cell 181 (2) (2020) 281–292.e6.

[57] J. Shang, Y. Wan, C. Luo, G. Ye, Q. Geng, A. Auerbach, F. Li, Cell entry mechanisms of SARS-CoV-2, Proc. Natl. Acad. Sci. USA 117 (21) (2020) 11727–11734.

[58] L. Ren, Y. Zhang, J. Li, Y. Xiao, J. Zhang, Y. Wang, L. Chen, G. Paranhos-Baccalà, J. Wang, Genetic drift of human coronavirus OC43 spike gene during adaptive evolution, Sci. Rep. 5 (1) (2015) 11451.

[59] C. Wang, Z. Liu, Z. Chen, X. Huang, M. Xu, T. He, Z. Zhang, The establishment of reference sequence for SARS-CoV-2 and variation analysis, J. Med. Virol. 92 (6) (2020) 667–674.

[60] The scikit-bio development team, Scikit-bio: A bioinformatics library for data scientists, students, and developers, 2020.

[61] M. Zhao, W.-P. Lee, E.P. Garrison, G.T. Marth, SSW library: An SIMD Smith-Waterman C/C++ Library for use in genomic applications, PLOS ONE 8 (12) (2013) e82138.

[62] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.

[63] National Institutes of Health, Clinical spectrum of SARS-CoV-2 infection, 2021.

[64] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, J. R. Stat. Soc. Ser. B Stat. Methodol. 67 (2) (2005) 301–320.

[65] P. Waldmann, G. Mészáros, B. Gredler, C. Fürst, J. Sölkner, Evaluation of the lasso and the elastic net in genome-wide association studies, Front. Genet. 4 (2013).

[66] N. Van Goethem, A. Robert, N. Bossuyt, L.A.E. Van Poelvoorde, S. Quoilin, S.C.J. De Keersmaecker, B. Devleesschauwer, I. Thomas, K. Vanneste, N.H.C. Roosens, H. Van Oyen, Evaluation of the added value of viral genomic information for predicting severity of influenza infection, BMC Infect. Dis. 21 (1) (2021) 785.

[67] J. Wang, M. Gribskov, IRESpy: An XGBoost model for prediction of internal ribosome entry sites, BMC Bioinformatics 20 (1) (2019) 409.

[68] T. ValizadehAslani, Z. Zhao, B.A. Sokhansanj, G.L. Rosen, Amino acid K-mer feature extraction for quantitative antimicrobial resistance (AMR) prediction by machine learning and model interpretation for biological insights, Biology 9 (11) (2020) E365.

[69] X. Liang, F. Li, J. Chen, J. Li, H. Wu, S. Li, J. Song, Q. Liu, Large-scale comparative review and assessment of computational methods for anti-cancer peptide identification, Brief. Bioinform. 22 (4) (2021) bbaa312.

[70] A.K. Benson, S.A. Kelly, R. Legge, F. Ma, S.J. Low, J. Kim, M. Zhang, P.L. Oh, D. Nehrenberg, K. Hua, S.D. Kachman, E.N. Moriyama, J. Walter, D.A. Peterson, D. Pomp, Individuality in gut microbiota composition is a complex polygenic trait shaped by multiple environmental and host genetic factors, Proc. Natl. Acad. Sci. USA 107 (44) (2010) 18933–18938.

[71] X. Zhang, B. Guo, N. Yi, Zero-inflated Gaussian mixed models for analyzing longitudinal microbiome data, PLoS ONE 15 (11) (2020) e0242073.

[72] Y. Jiang, J. Chen, W. Chen, Controlling batch effect in epigenome-wide association study, Methods Mol. Biol. (Clifton, N.J.) 2432 (2022) 73–84.

[73] C. Ngufor, H. Van Houten, B.S. Caffo, N.D. Shah, R.G. McCoy, Mixed effect machine learning: A framework for predicting longitudinal change in hemoglobin A1c, J. Biomed. Inform. 89 (2019) 56–67.

[74] F. Zhou, A. Alsaid, M. Blommer, R. Curry, R. Swaminathan, D. Kochhar, W. Talamonti, L. Tijerina, Predicting driver fatigue in monotonous automated driving with explanation using gpboost and SHAP, Int. J. Human Comput. Interact. 38 (8) (2022) 719–729.

[75] S. Ramraj, N. Uzir, R. Sunil, S. Banerjee, Experimenting XGBoost algorithm for prediction and classification of different datasets, Int. J. Control Theory Appl. 9 (2016) 651–662.

[76] J. Elith, J.R. Leathwick, T. Hastie, A working guide to boosted regression trees, J. Anim. Ecol. 77 (4) (2008) 802–813.

[77] G. Grasselli, M. Greco, A. Zanella, G. Albano, M. Antonelli, G. Bellani, E. Bonanomi, L. Cabrini, E. Carlesso, G. Castelli, S. Cattaneo, D. Cereda, S. Colombo, A. Coluccello, G. Crescini, A. Forastieri Molinari, G. Foti, R. Fumagalli, G.A. Iotti, T. Langer, N. Latronico, F.L. Lorini, F. Mojoli, G. Natalini, C.M. Pessina, V.M. Ranieri, R. Rech, L. Scudeller, A. Rosano, E. Storti, B.T. Thompson, M. Tirani, P.G. Villani, A. Pesenti, M. Cecconi, COVID-19 Lombardy ICU Network, Risk factors associated with mortality among patients with COVID-19 in intensive care units in Lombardy, Italy, JAMA Internal Med. 180 (10) (2020) 1345–1355.

[78] H. Holt, M. Talaei, M. Greenig, D. Zenner, J. Symons, C. Relton, K.S. Young, M.R. Davies, K.N. Thompson, J. Ashman, S.S. Rajpoot, A.A. Kayyale, S. El Rifai, P.J. Lloyd, D. Jolliffe, O. Timmis, S. Finer, S. Iliodromiti, A. Miners, N.S. Hopkinson, B. Alam, G. Lloyd-Jones, T. Dietrich, I. Chapple, P.E. Pfeffer, D. McCoy, G. Davies, R.A. Lyons, C. Griffiths, F. Kee, A. Sheikh, G. Breen, S.O. Shaheen, A.R. Martineau, Risk factors for developing COVID-19: A population-based longitudinal study (COVIDENCE UK), Thorax (2021) thoraxjnl–2021–217487.

[79] H. Peckham, N.M. de Gruijter, C. Raine, A. Radziszewska, C. Ciurtin, L.R. Wedderburn, E.C. Rosser, K. Webb, C.T. Deakin, Male sex identified by global COVID-19 meta-analysis as a risk factor for death and ITU admission, Nature Commun. 11 (1) (2020) 6317.

[80] S. Mukherjee, K. Pahan, Is COVID-19 gender-sensitive? J. Neuroimmune Pharmacol.: Off. J. Soc. NeuroImmune Pharmacol. 16 (1) (2021) 38–47.

[81] S.H. Hsu, S.-H. Chang, C.P. Gross, S.-Y. Wang, Relative risks of COVID-19 fatality between the first and second waves of the pandemic in Ontario, Canada, Int. J. Infect. Dis.: IJID : Off. Publ. Int. Soc. Infect. Dis. 109 (2021) 189–191.

[82] J. Lopez Bernal, N. Andrews, C. Gower, C. Robertson, J. Stowe, E. Tessier, R. Simmons, S. Cottrell, R. Roberts, M. O'Doherty, K. Brown, C. Cameron, D. Stockton, J. McMenamin, M. Ramsay, Effectiveness of the Pfizer-BioNTech and Oxford-AstraZeneca vaccines on Covid-19 related symptoms, hospital admissions, and mortality in older adults in England: Test negative case-control study, BMJ (Clin. Res. Ed.) 373 (2021) n1088.

[83] T. Akpolat, O. Uzun, Reduced mortality rate after coronavac vaccine among healthcare workers, J. Infect. 83 (2) (2021) e20–e21.

[84] E.J. Haas, F.J. Angulo, J.M. McLaughlin, E. Anis, S.R. Singer, F. Khan, N. Brooks, M. Smaja, G. Mircus, K. Pan, J. Southern, D.L. Swerdlow, L. Jodar, Y. Levy, S. Alroy-Preis, Impact and effectiveness of mRNA BNT162b2 vaccine against SARS-CoV-2 infections and COVID-19 cases, hospitalisations, and deaths following a nationwide vaccination campaign in Israel: An observational study using national surveillance data, Lancet (London, England) 397 (10287) (2021) 1819–1829.

[85] A.A. Grima, K.R. Murison, A.E. Simmons, A.R. Tuite, D.N. Fisman, Relative virulence of SARS-CoV-2 among vaccinated and unvaccinated individuals hospitalized with SARS-CoV-2, Clin. Infect. Dis.: Off. Publ. Infect. Dis. Soc. Am. (2022) ciac412.
[86] N.R. Aggarwal, L.E. Beaty, T.D. Bennett, N.E. Carlson, C.B. Davis, B.M. Kwan, D.A. Mayer, T.C. Ong, S. Russell, J. Steele, A.F. Wogu, M.K. Wynia, R.D. Zane, A.A. Ginde, Real world evidence of the neutralizing monoclonal antibody sotrovimab for preventing hospitalization and mortality in COVID-19 outpatients, MedRxiv: Prepr. Serv. Health Sci. (2022) 2022.04.03.22273360.
[87] G. Onder, G. Rezza, S. Brusaferro, Case-fatality rate and characteristics of patients dying in relation to COVID-19 in Italy, JAMA 323 (18) (2020) 1775–1776.
[88] S. Mahajan, C. Caraballo, S.-X. Li, Y. Dong, L. Chen, S.K. Huston, R. Srinivasan, C.A. Redlich, A.I. Ko, J.S. Faust, H.P. Forman, H.M. Krumholz, SARS-CoV-2 infection hospitalization rate and infection fatality rate among the non-congregate population in connecticut, Am. J. Med. 134 (6) (2021) 812–816.e2.
[89] W. Yang, S. Kandula, M. Huynh, S.K. Greene, G. Van Wye, W. Li, H.T. Chan, E. McGibbon, A. Yeung, D. Olson, A. Fine, J. Shaman, Estimating the infection-fatality risk of SARS-CoV-2 in New York City during the spring 2020 pandemic wave: A model-based analysis, Lancet Infect. Dis. 21 (2) (2021) 203–212.
[90] Z. Zhao, B.A. Sokhansanj, C. Malhotra, K. Zheng, G.L. Rosen, Genetic grouping of SARS-CoV-2 coronavirus sequences using informative subtype markers for pandemic spread visualization, PLoS Comput. Biol. 16 (9) (2020) e1008269.
[91] S.S. Negi, C.H. Schein, W. Braun, Regional and temporal coordinated mutation patterns in SARS-CoV-2 spike protein revealed by a clustering and network analysis, Sci. Rep. 12 (1) (2022) 1128.
[92] M. Monod, A. Blenkinsop, X. Xi, D. Hebert, S. Bershan, S. Tietze, M. Baguelin, V.C. Bradley, Y. Chen, H. Coupland, S. Filippi, J. Ish-Horowicz, M. McManus, T. Mellan, A. Gandy, M. Hutchinson, H.J.T. Unwin, S.L. van Elsland, M.A.C. Vollmer, S. Weber, H. Zhu, A. Bezancon, N.M. Ferguson, S. Mishra, S. Flaxman, S. Bhatt, O. Ratmann, Age groups that sustain resurging COVID-19 epidemics in the United States, Science 371 (6536) (2021).
[93] M.R. Islam, M.N. Hoque, M.S. Rahman, A.S.M.R.U. Alam, M. Akther, J.A. Puspo, S. Akter, M. Sultana, K.A. Crandall, M.A. Hossain, Genome-wide analysis of SARS-CoV-2 virus strains circulating worldwide implicates heterogeneity, Sci. Rep. 10 (1) (2020) 14004.
[94] Z. Chen, K.C. Chong, M.C.S. Wong, S.S. Boon, J. Huang, M.H. Wang, R.W.Y. Ng, C.K.C. Lai, P.K.S. Chan, A global analysis of replacement of genetic variants of SARS-CoV-2 in association with containment capacity and changes in disease severity, Clin. Microbiol. Infect.: Off. Publ. Eur. Soc. Clin. Microbiol. Infect. Dis. 27 (5) (2021) 750–757.
[95] A.L. Oberg, D.W. Mahoney, Linear mixed effects models, Methods Mol. Biol. (Clifton, N.J.) 404 (2007) 213–234.
[96] I. Lazarevic, V. Pravica, D. Miljanovic, M. Cupic, Immune evasion of SARS-CoV-2 emerging variants: What have we learnt so far?, Viruses 13 (7) (2021) 1192.
[97] M. Noori, S.A. Nejadghaderi, S. Arshi, K. Carson-Chahhoud, K. Ansarin, A.-A. Kolahi, S. Safiri, Potency of BNT162b2 and mRNA-1273 vaccine-induced neutralizing antibodies against severe acute respiratory syndrome-CoV-2 variants of concern: A systematic review of in vitro studies, Rev. Med. Virol. 32 (2) (2022) e2277.
[98] C.K.V. Nonaka, T. Gräf, C.A.d.L. Barcia, V.F. Costa, J.L. de Oliveira, R.d.H. Passos, I.N. Bastos, M.C.B. de Santana, I.M. Santos, K.A.F. de Sousa, T.G.L. Weber, I.C. de Siqueira, C.A.G. Rocha, A.V.A. Mendes, B.S.d.F. Souza, SARS-CoV-2 variant of concern P.1 (Gamma) infection in Young and middle-aged patients admitted to the intensive care units of a single hospital in Salvador, Northeast Brazil, February 2021, Int. J. Infect. Dis. 111 (2021) 47–54.
[99] S. Albaradei, M. Thafar, A. Alsaedi, C. Van Neste, T. Gojobori, M. Essack, X. Gao, Machine learning and deep learning methods that use omics data for metastasis prediction, Comput. Struct. Biotechnol. J. 19 (2021) 5008–5018.
[100] P. Domingos, A few useful things to know about machine learning, Commun. ACM 55 (10) (2012) 78–87.
[101] M. Dhawan, A. Sharma, n. Priyanka, N. Thakur, T.K. Rajkhowa, O.P. Choudhary, Delta variant (B.1.617.2) of SARS-CoV-2: Mutations, impact, challenges and possible solutions, Human Vaccines Immunother. (2022) 2068883.
[102] A. Saito, T. Irie, R. Suzuki, T. Maemura, H. Nasser, K. Uriu, Y. Kosugi, K. Shirakawa, K. Sadamasu, I. Kimura, J. Ito, J. Wu, K. Iwatsuki-Horimoto, M. Ito, S. Yamayoshi, S. Ozono, E.P. Butlertanaka, Y.L. Tanaka, R. Shimizu, K. Shimizu, K. Yoshimatsu, R. Kawabata, T. Sakaguchi, K. Tokunaga, I. Yoshida, H. Asakura, M. Nagashima, Y. Kazuma, R. Nomura, Y. Horisawa, K. Yoshimura, A. Takaori-Kondo, M. Imai, S. Nakagawa, T. Ikeda, T. Fukuhara, Y. Kawaoka, K. Sato, The Genotype to Phenotype Japan (G2P-Japan) Consortium, SARS-CoV-2 spike P681R mutation, a hallmark of the Delta variant, enhances viral fusogenicity and pathogenicity, 2021, http://dx.doi.org/10.1101/2021.06.17.448820.
[103] A. Kuzmina, N. Atari, A. Ottolenghi, D. Korovin, I.C. Lass, B. Rosental, E. Rosenberg, M. Mandelboim, R. Taube, P681 mutations within the polybasic motif of spike dictate fusogenicity and syncytia formation of SARS CoV-2 variants, 2022, http://dx.doi.org/10.1101/2022.04.26.489630.
[104] Y. Zhang, T. Zhang, Y. Fang, J. Liu, Q. Ye, L. Ding, SARS-CoV-2 spike L452R mutation increases Omicron variant fusogenicity and infectivity as well as host glycolysis, Signal Transduct. Target. Ther. 7 (1) (2022) 1–3.
[105] C. Motozono, M. Toyoda, J. Zahradnik, A. Saito, H. Nasser, T.S. Tan, I. Ngare, I. Kimura, K. Uriu, Y. Kosugi, Y. Yue, R. Shimizu, J. Ito, S. Torii, A. Yonekawa, N. Shimono, Y. Nagasaki, R. Minami, T. Toya, N. Sekiya, T. Fukuhara, Y. Matsuura, G. Schreiber, T. Ikeda, S. Nakagawa, T. Ueno, K. Sato, Genotype to Phenotype Japan (G2P-Japan) Consortium, SARS-CoV-2 spike L452R variant evades cellular immunity and increases infectivity, Cell Host Microbe 29 (7) (2021) 1124–1136.e11.
[106] K. Bansal, S. Kumar, Mutational cascade of SARS-CoV-2 leading to evolution and emergence of omicron variant, Virus Res. 315 (2022) 198765.
[107] L. Schnirring, ECDC ups BA.4, BA.5 to variants of concern, warns of case rises, CIDRAP (2022).
[108] A. Maxmen, Why call it BA.2.12.1? A guide to the tangled Omicron family, Nature (2022).
[109] R. Uraki, M. Kiso, S. Iida, M. Imai, E. Takashita, M. Kuroda, P.J. Halfmann, S. Loeber, T. Maemura, S. Yamayoshi, S. Fujisaki, Z. Wang, M. Ito, M. Ujie, K. Iwatsuki-Horimoto, Y. Furusawa, R. Wright, Z. Chong, S. Ozono, A. Yasuhara, H. Ueki, Y. Sakai-Tagawa, R. Li, Y. Liu, D. Larson, M. Koga, T. Tsutsumi, E. Adachi, M. Saito, S. Yamamoto, M. Hagihara, K. Mitamura, T. Sato, M. Hojo, S.-I. Hattori, K. Maeda, R. Valdez, M. Okuda, J. Murakami, C. Duong, S. Godbole, D.C. Douek, K. Maeda, S. Watanabe, A. Gordon, N. Ohmagari, H. Yotsuyanagi, M.S. Diamond, H. Hasegawa, H. Mitsuya, T. Suzuki, Y. Kawaoka, IASO study team, Characterization and antiviral susceptibility of SARS-CoV-2 omicron/BA.2, Nature (2022).
[110] D. Yamasoba, I. Kimura, H. Nasser, Y. Morioka, N. Nao, J. Ito, K. Uriu, M. Tsuda, J. Zahradnik, K. Shirakawa, R. Suzuki, M. Kishimoto, Y. Kosugi, K. Kobiyama, T. Hara, M. Toyoda, Y.L. Tanaka, E.P. Butlertanaka, R. Shimizu, H. Ito, L. Wang, Y. Oda, Y. Orba, M. Sasaki, K. Nagata, K. Yoshimatsu, H. Asakura, M. Nagashima, K. Sadamasu, K. Yoshimura, J. Kuramochi, M. Seki, R. Fujiki, A. Kaneda, T. Shimada, T.-a. Nakada, S. Sakao, T. Suzuki, T. Ueno, A. Takaori-Kondo, K.J. Ishii, G. Schreiber, H. Sawa, A. Saito, T. Irie, S. Tanaka, K. Matsuno, T. Fukuhara, T. Ikeda, K. Sato, The Genotype to Phenotype Japan (G2P-Japan) Consortium, Virological characteristics of SARS-CoV-2 BA.2 variant, 2022, http://dx.doi.org/10.1101/2022.02.14.480335.
[111] M. Whitaker, J. Elliott, B. Bodinier, W. Barclay, H. Ward, G. Cooke, C.A. Donnelly, M. Chadeau-Hyam, P. Elliott, Variant-specific symptoms of COVID-19 among 1,542,510 people in England, 2022, http://dx.doi.org/10.1101/2022.05.21.22275368.
[112] D. Loconsole, F. Centrone, A. Sallustio, M. Accogli, D. Casulli, D. Sacco, R. Zagaria, C. Morcavallo, M. Chironna, Characteristics of the first 284 patients infected with the SARS-CoV-2 omicron BA.2 subvariant at a single center in the apulia region of Italy, January–March 2022, Vaccines 10 (5) (2022) 674.
[113] J. Yu, A.-r.Y. Collier, M. Rowe, F. Mardas, J.D. Ventura, H. Wan, J. Miller, O. Powers, B. Chung, M. Siamatu, N.P. Hachmann, N. Surve, F. Nampanya, A. Chandrashekar, D.H. Barouch, Neutralization of the SARS-CoV-2 omicron BA.1 and BA.2 variants, N. Engl. J. Med. 386 (16) (2022) 1579–1580.
[114] L. Liu, S. Iketani, Y. Guo, J.F.-W. Chan, M. Wang, L. Liu, Y. Luo, H. Chu, Y. Huang, M.S. Nair, J. Yu, K.K.-H. Chik, T.T.-T. Yuen, C. Yoon, K.K.-W. To, H. Chen, M.T. Yin, M.E. Sobieszczyk, Y. Huang, H.H. Wang, Z. Sheng, K.-Y. Yuen, D.D. Ho, Striking antibody evasion manifested by the Omicron variant of SARS-CoV-2, Nature 602 (7898) (2022) 676–681.
[115] S. Iketani, L. Liu, Y. Guo, L. Liu, J.F.-W. Chan, Y. Huang, M. Wang, Y. Luo, J. Yu, H. Chu, K.K.-H. Chik, T.T.-T. Yuen, M.T. Yin, M.E. Sobieszczyk, Y. Huang, K.-Y. Yuen, H.H. Wang, Z. Sheng, D.D. Ho, Antibody evasion properties of SARS-CoV-2 Omicron sublineages, Nature 604 (7906) (2022) 553–556.
[116] A.-C.S. Vogt, G. Augusto, B. Martina, X. Chang, G. Nasrallah, D.E. Speiser, M. Vogel, M.F. Bachmann, M.O. Mohsen, Increased receptor affinity and reduced recognition by specific antibodies contribute to immune escape of SARS-CoV-2 variant omicron, Vaccines 10 (5) (2022) 743.
[117] J. Quandt, A. Muik, N. Salisch, B.G. Lui, S. Lutz, K. Krüger, A.-K. Wallisch, P. Adams-Quack, M. Bacher, A. Finlayson, O. Ozhelvaci, I. Vogler, K. Grikscheit, S. Hoehl, U. Goetsch, S. Ciesek, O. Türeci, U. Sahin, Omicron BA.1 breakthrough infection drives cross-variant neutralization and memory B cell formation against conserved epitopes, Sci. Immunol. (2022) eabq2427.
[118] Q. Wang, Y. Guo, S. Iketani, Z. Li, H. Mohri, M. Wang, J. Yu, A.D. Bowen, J.Y. Chang, J.G. Shah, N. Nguyen, K. Meyers, M.T. Yin, M.E. Sobieszczyk, Z. Sheng, Y. Huang, L. Liu, D.D. Ho, SARS-CoV-2 omicron BA.2.12.1, BA.4, and BA.5 subvariants evolved to extend antibody evasion, 2022, http://dx.doi.org/10.1101/2022.05.26.493517.
[119] C.N. Andreassen, A simulated SNP experiment indicates a high risk of over-fitting and false positive results when a predictive multiple SNP model is established and tested within the same dataset, Radiother. Oncol.: J. Eur. Soc. Ther. Radiol. Oncol. 114 (3) (2015) 310–313.
[120] D.T. Jones, Setting the standards for machine learning in biology, Nat. Rev. Mol. Cell Biol. 20 (11) (2019) 659–660.
[121] Y. Takahashi, M. Ueki, G. Tamiya, S. Ogishima, K. Kinoshita, A. Hozawa, N. Minegishi, F. Nagami, K. Fukumoto, K. Otsuka, K. Tanno, K. Sakata, A. Shimizu, M. Sasaki, K. Sobue, S. Kure, M. Yamamoto, H. Tomita, Machine learning for effectively avoiding overfitting is a crucial strategy for the genetic prediction of polygenic psychiatric phenotypes, Transl. Psychiatry 10 (1) (2020) 1–11.

[122] N. Mikolajewicz, S.V. Komarova, Meta-analytic methodology for basic research: A practical guide, Front. Physiol. 10 (2019).
[123] L.M. Schriml, M. Chuvochina, N. Davies, E.A. Eloe-Fadrosh, R.D. Finn, P. Hugenholtz, C.I. Hunter, B.L. Hurwitz, N.C. Kyrpides, F. Meyer, I.K. Mizrachi, S.-A. Sansone, G. Sutton, S. Tighe, R. Walls, COVID-19 pandemic reveals the peril of ignoring metadata standards, Sci. Data 7 (1) (2020) 188.
[124] R.P. Bhattacharyya, W.P. Hanage, Challenges in inferring intrinsic severity of the SARS-CoV-2 omicron variant, N. Engl. J. Med. 386 (7) (2022) e14.
[125] M.S. Calderwood, V.M. Deloney, D.J. Anderson, V.C.-C. Cheng, S. Gohil, J.H. Kwon, L. Mody, E. Monsees, V.M. Vaughn, T.L. Wiemken, M.J. Ziegler, E. Lofgren, Policies and practices of SHEA research network hospitals during the COVID-19 pandemic, Infect. Control Hosp. Epidemiol. 41 (10) (2020) 1127–1135.
[126] N. Fillmore, J. La, C. Zheng, S. Doron, N. Do, P. Monach, W. Branch-Elliman, The COVID-19 Hospitalization Metric in the Pre- and Post-Vaccination Eras as a Measure of Pandemic Severity: A Retrospective, Nationwide Cohort Study, Preprint, 2021, http://dx.doi.org/10.21203/rs.3.rs-898254/v1, In Review.
[127] L.E. Kushner, A.R. Schroeder, J. Kim, R. Mathew, "For COVID" or "with COVID": Classification of SARS-CoV-2 hospitalizations in children, Hosp. Pediatr. 11 (8) (2021) e151–e156.
[128] N.E. Webb, T.S. Osburn, Characteristics of hospitalized children positive for SARS-CoV-2: Experience of a large center, Hosp. Pediatr. 11 (8) (2021) e133–e141.
[129] L. Rocheleau, G. Laroche, K. Fu, C.M. Stewart, A.O. Mohamud, M. Côté, P.M. Giguère, M.-A. Langlois, M. Pelchat, Identification of a high-frequency intrahost SARS-CoV-2 spike variant with enhanced cytopathic and fusogenic effects, MBio (2021).
[130] D. Jacot, T. Pillonel, G. Greub, C. Bertelli, Assessment of SARS-CoV-2 genome sequencing: Quality criteria and low-frequency variants, J. Clin. Microbiol. 59 (10) (2021) e0094421.
[131] K.A. Lagerborg, E. Normandin, M.R. Bauer, G. Adams, K. Figueroa, C. Loreth, A. Gladden-Young, B.M. Shaw, L.R. Pearlman, D. Berenzy, H.B. Dewey, S. Kales, S.T. Dobbins, E.S. Shenoy, D. Hooper, V.M. Pierce, K.C. Zachary, D.J. Park, B.L. MacInnis, R. Tewhey, J.E. Lemieux, P.C. Sabeti, S.K. Reilly, K.J. Siddle, Synthetic DNA spike-ins (SDSIs) enable sample tracking and detection of inter-sample contamination in SARS-CoV-2 sequencing workflows, Nat. Microbiol. 7 (1) (2022) 108–119.
[132] H. Ejaz, A. Alsrhani, A. Zafar, H. Javed, K. Junaid, A.E. Abdalla, K.O.A. Abosalif, Z. Ahmed, S. Younas, COVID-19 and comorbidities: Deleterious impact on infected patients, J. Infect. Public Health 13 (12) (2020) 1833–1839.
[133] Z.G. Dessie, T. Zewotir, Mortality-related risk factors of COVID-19: A systematic review and meta-analysis of 42 studies and 423,117 patients, BMC Infect. Dis. 21 (1) (2021) 855.
[134] S.-W. Huang, S.-F. Wang, SARS-CoV-2 entry related viral and host genetic variations: Implications on COVID-19 severity, immune escape, and infectivity, Int. J. Mol. Sci. 22 (6) (2021) 3060.
[135] S. Mohammadpour, A. Torshizi Esfahani, M. Halaji, M. Lak, R. Ranjbar, An updated review of the association of host genetic factors with susceptibility and resistance to COVID-19, J. Cell. Physiol. 236 (1) (2021) 49–54.
[136] I. Fricke-Galindo, R. Falfán-Valencia, Genetics insight for COVID-19 susceptibility and severity: A review, Front. Immunol. 12 (2021) 622176.
[137] A. Verma, N.L. Tsao, L.O. Thomann, Y.-L. Ho, S.K. Iyengar, S.-W. Luoh, R. Carr, D.C. Crawford, J.T. Efird, J.E. Huffman, A. Hung, K.L. Ivey, M.G. Levin, J. Lynch, P. Natarajan, S. Pyarajan, A.G. Bick, L. Costa, G. Genovese, R. Hauger, R. Madduri, G.A. Pathak, R. Polimanti, B. Voight, M. Vujkovic, S.M. Zekavat, H. Zhao, M.D. Ritchie, K.-M. Chang, K. Cho, J.P. Casas, P.S. Tsao, J.M. Gaziano, C. O'Donnell, S.M. Damrauer, K.P. Liao, VA Million Veteran Program COVID-19 Science Initiative, A phenome-wide association study of genes associated with COVID-19 severity reveals shared genetics with complex diseases in the million veteran program, PLOS Genet. 18 (4) (2022) e1010113.
[138] S. Chlamydas, A.G. Papavassiliou, C. Piperi, Epigenetic mechanisms regulating COVID-19 infection, Epigenetics 16 (3) (2021) 263–270.
[139] A.B.M.M.K. Islam, M.A.-A.-K. Khan, R. Ahmed, M.S. Hossain, S.M.T. Kabir, M.S. Islam, A.M.A.M.Z. Siddiki, Transcriptome of nasopharyngeal samples from COVID-19 patients and a comparative analysis with other SARS-CoV-2 infection models reveal disparate host responses against SARS-CoV-2, J. Transl. Med. 19 (2021) 32.
[140] H. Zhao, A. Nguyen, D. Wu, Y. Li, S.A. Hassan, J. Chen, H. Shroff, G. Piszczek, P. Schuck, Plasticity in structure and assembly of SARS-CoV-2 nucleocapsid protein, PNAS Nexus (2022) pgac049.
[141] T.N. Starr, A.J. Greaney, A. Addetia, W.W. Hannon, M.C. Choudhary, A.S. Dingens, J.Z. Li, J.D. Bloom, Prospective mapping of viral mutations that escape antibodies used to treat COVID-19, Science 371 (6531) (2021) 850–854.
[142] M. Puray-Chavez, K.M. LaPak, T.P. Schrank, J.L. Elliott, D.P. Bhatt, M.J. Agajanian, R. Jasuja, D.Q. Lawson, K. Davis, P.W. Rothlauf, Z. Liu, H. Jo, N. Lee, K. Tenneti, J.E. Eschbach, C.S. Mugisha, E.M. Cousins, E.W. Cloer, H.R. Vuong, L.A. VanBlargan, A.L. Bailey, P. Gilchuk, J.E. Crowe, M.S. Diamond, D.N. Hayes, S.P.J. Whelan, A. Horani, S.L. Brody, D. Goldfarb, M.B. Major, S.B. Kutluay, Systematic analysis of SARS-CoV-2 infection of an ACE2-negative human airway cell, Cell Rep. 36 (2) (2021).
[143] A.J. Greaney, T.N. Starr, P. Gilchuk, S.J. Zost, E. Binshtein, A.N. Loes, S.K. Hilton, J. Huddleston, R. Eguia, K.H. Crawford, A.S. Dingens, R.S. Nargi, R.E. Sutton, N. Suryadevara, P.W. Rothlauf, Z. Liu, S.P. Whelan, R.H. Carnahan, J.E. Crowe, J.D. Bloom, Complete mapping of mutations to the SARS-CoV-2 spike receptor-binding domain that escape antibody recognition, Cell Host Microbe 29 (1) (2021) 44–57.e9.
[144] M. Torrens-Fontanals, A. Peralta-García, C. Talarico, R. Guixà-González, T. Giorgino, J. Selent, SCoV2-MD: A database for the dynamics of the SARS-CoV-2 proteome and variant impact predictions, Nucleic Acids Res. 50 (D1) (2022) D858–D866.
[145] A. Kaur, A.S. Chauhan, A. kumar Aggarwal, Prediction of enhancers in DNA sequence data using a hybrid CNN-DLSTM model, IEEE/ACM Trans. Comput. Biol. Bioinform. (2022) 1.
[146] A. Kaur, A.S. Chauhan, A. kumar Aggarwal, Dynamic deep genomics sequence encoder for managed file transfer, IETE J. Res. (2022).
[147] M.L. Bileschi, D. Belanger, D.H. Bryant, T. Sanderson, B. Carter, D. Sculley, A. Bateman, M.A. DePristo, L.J. Colwell, Using deep learning to annotate the protein universe, Nature Biotechnol. (2022).