Characterizing and Understanding Temporal Effects in COVID-19 Data

Bruno Barbosa Miranda de Paiva [1] **Polianna Delfino-Pereira** [2] **Virginia Mara Reis Gomes** [2]
Maira Viana Rego Souza-Silva [2] **Cláudio Valiense** [1] **Milena Soriano Marcolino** [2] **Marcos André Gonçalves** [1]

Abstract

Since the global outbreak of the coronavirus 2019 pandemic, hundreds of works have been published, analyzing and modeling multiple aspects of the disease. Several of them venture into predictive and modeling tasks, such as mortality prediction and patient severity scoring, using machine-learning (ML) algorithms. An important limitation for most of these works is the fact that they do not consider the multiple temporal aspects of this pandemic, especially regarding disease profile and distributional changes over the months. Such temporal effects are mostly due to multiple interactions between different and novel viral strains, combined with mass vaccination campaigns targeting different groups or patterns (e.g., prioritizing older individuals and those with comorbidity first) and availability of different vaccines. These temporal effects result in impaired model effectiveness and classification errors. In this paper, using a large dataset with over 10,000 patients from 39 hospitals in Brazil admitted during a period of more than 20 months, we provide an overview of the multiple forms of temporal drift that happened during the pandemic and the magnitude of their effects on model effectiveness. Our analyses encompass changes in the severely ill patients' profile as well as how mortality rates have changed over time. We also investigate how the importance of different predictive variables change and shift over time.

1. Introduction

Ever since the coronavirus 2019 pandemic outbreak, the number of cases and deaths has increased exponentially. As of May 2022, over 500 million cases and 6 million deaths have been officially reported. Although vaccines have been developed and produced at unprecedented speeds, they have been administered at a slow and uneven roll out, what occurred concurrently with the emergence of new variants. Indeed, the number of people reinfected with the Omicron and other variants of SARS-CoV-2 increased sharply, despite the vaccines' robust protection against serious illness, hospitalization or death. Accordingly, many pandemic's aspects have changed over time.

In this context, various modeling tools have been developed to assist effective decision-making. A common sub task in this kind of work is early patient risk stratification for predicting inhospital COVID-19 mortality. As suggested by Marcolino et al. (Marcolino et al., 2021), many approaches have been used, such as those described by (Ikemura et al., 2021; Paiva et al., 2022), which tested different machine learning algorithms to find high-effectiveness models to predict the mortality risk of COVID-19 patients (Ikemura et al., 2021; Paiva et al., 2022). A key limitation in most of these works is disregarding the impact of temporal data drifts while using past data to learn for both classification and regression tasks.

Temporal data aspects may yield a significant impact on the final model effectiveness at predictive tasks, as different kinds of feature and label drifts become "training noise". Such drifts may cause further shifting of the classification boundaries into improper regions. For instance, vaccination strategies prioritized the elderly first, shifting the overall age of dying patients towards the younger population, while also altering potential casualty rates and interacting with other time-related features. Concurrently, the dynamic interactions between changes in population immunity, ongoing viral evolution and immune escape have driven the spread of viral variants, with different transmissibility and disease severity profile (Telenti et al., 2021).

Some works have been proposed with the intention of characterizing differences between various waves of the pandemic, such as (Zeiser et al., 2022), (Iftimie et al., 2021) and (Carbonell et al., 2021). These works have consistently shown differences in mortality and hospitalization profiles between waves, as measured in different countries and cities. However, none of them have *characterized and*

[1]Department of Computer Science, University of Minas Gerais, Belo Horizonte, Brazil [2]Department of Internal Medicine, University of Minas Gerais, Belo Horizonte, Brazil. Correspondence to: Bruno Paiva <brunobarbosa.mpaiva@gmail.com>.

Proceedings of the 1st Workshop on Healthcare AI and COVID-19, ICML 2022.

quantified such differences from a data drift perspective, which is the main contribution of this paper.

Temporal drifts are known phenomena in other contexts and related characterizations has been performed before in works such as (Mourao et al., 2008) and (Salles et al., 2016), which have analyzed drifts in word distributions, word meanings and target class distributions. However in the context of COVID-19, we are unaware of any work that provides such a temporal characterization of data/concept drifts and their potential in classification and predictive tasks.

Summarizing, in this paper, we aim at identifying and characterizing different types of temporal drifts using past COVID-19 data. For this, we exploit a large dataset obtained through the Brazilian Multicenter COVID-19 Registry, with more than 10,000 in-hospital patients from 39 hospitals, who were admitted during a period of more than 20 months. We provide an overview of the multiple forms of temporal drift that happened during the pandemic and the magnitude of their effects on model effectiveness. Our analyses encompass changes in the severely ill patients' profiles as well as changes in mortality rates over time. Furthermore, we investigate how the importance of the most predictive features (variables) shift over time due to analyzed temporal data drifts. Finally, We observed significant drops in classifier effectiveness over time, that can be potentially explained by the observed temporal drifts.

This work is organized as follows. Section 2 discusses related work. We proceed with the characterization of the temporal effects in our COVID-19 dataset. We finish with conclusions and perspectives for future work.

2. Related Work

There have been multiple studies analyzing variations observed over time in the class distribution. Studies such as (Salles et al., 2016) and (Mourao et al., 2008), for instance, perform a detailed characterization of such effects in textual datasets of documents organized into topics. The characterization of these effects with regards to the COVID-19 pandemic, however, is a scarcely studied problem. Indeed, we were able to find only two other works (Jung et al., 2022; Jassat et al., 2021) showing differences in hospitalized patient profiles as new COVID-19 waves began spreading. (Jung et al., 2022), for instance, used multivariate logistic regression and a complementary machine learning-based analysis using explainability methods for better understanding the influence of age and comorbidities during the different pandemic waves. The authors observed that an older age was not as important risk factor to develop severe COVID-19 in the third wave as it was in the previous one. Meanwhile, (Jassat et al., 2021) used multivariate logistic regression to investigate the increase in adjusted

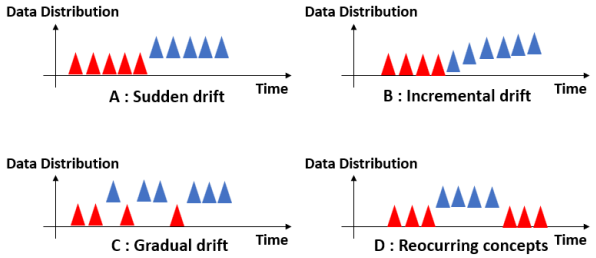

Figure 1. Drift types with respect to the passing of time.

mortality when comparing the second-wave to the first one in South Africa, and attributed this variation to a new COVID-19 variant. We, on a complementary note, propose to identify and characterize different types of temporal drifts in a large Brazilian COVID-19 Registry. Such study has a large potential to help in the allocation of medical resources needed to manage patients in hospitals, and may help to define target groups for vaccine booster shots.

In more details, concept drifts describe changes in the statistical properties of target variables, while data drifts refer to input distribution changes (i.e. changes in target class distributions). These notions have been well defined in studies, such as (Moreno-Torres et al., 2012), which unified and consolidated some of the underlying terminologies. As defined by (Gama & Zhang, 2019), data and concept drifts can be categorized with regards to how they behave with regard to the passing of time, being: (i) sudden (i.e. one event permanently changes the "meaning" of a concept), (ii) incremental (i.e. one event incrementally generates gradual changes to the "meaning" of a concept); (iii) gradual (i.e. the concepts interchange gradually until the complete shift occurs), or (iv) reoccurring (i.e. a transient concept drift). These concepts are exemplified in Figure 1.

Approaches do exist to detect and learn in the presence of concept drifts. In most contexts, naively monitoring data drifts may be an expensive task, as it could require data labeling. As an alternative approach, (Haque et al., 2016) uses an ensemble of classifiers to report their prediction confidences and monitor changes in their confidence distribution, to detect the moment at which concept drift occurred. In our specific dataset and task, however, deaths are easily obtainable labeled data, which means that our main issue would be related to learning in the presence of data drift. In spite of that, different COVID-19 problems may also require data drift monitoring, such as COVID-19 detection on imaging tests.

On the issue of learning in the presence of data drifts, some solutions also have been reported in the literature, mostly focused on sample selection and/or sample weighting with variations on how they derive the final weighting and/or sampling. (Klinkenberg, 2004), for instance, tackles the problem

by using support vector machines (SVMs) for both sample selection and sample weighting through an iterative process that sequentially trains SVMs to find the training instances that constitutes the model's support vectors, weighting them based on how novel they are with respect to the desired timeline. (Kolter & Maloof, 2007) uses an special weighted ensemble to learn in the presence of such drifts. Other works such as (Salles et al., 2016; 2010) use a temporal weighting function that can be automatically learned (Salles et al., 2017) to select relevant samples for each training window.

Another interesting concept proposed by (Rocha et al., 2008) to tackle the problem is that of temporal contexts. In their work, which analyzes document collections, the authors define a temporal context as portions of documents that minimize the temporal effects of class distribution, term distribution and class similarity. This idea is used to devise a greedy strategy to optimize the trade-off between undersampling and temporal effects.

3. Dataset

This study exploits a dataset from a multicenter cohort, which included 10,897 adult Brazilian COVID-19 patients, admitted from March 2020 to November 2021 at 39 hospitals. This is a proprietary dataset, obtained through the course of said cohort. Data can be made available upon reasonable request. The median age of patients was 60 years-old [interquartile range 48-71] and 46% of them were women. In total, 21% of all registered patients died, yielding an unbalanced classification problem when predicting future deaths. The dataset consists of over 200 features, but only about 45 were admission features used in our future death classification task. Admission features in the dataset consist of data including patient's age, sex, comorbidities, laboratory tests (such as haemoglobin, C-reactive protein and leukocytes) and vital signs at hospital presentation (i.e. arterial blood pressure, respiratory rate, heart rate, etc).

Table 1 presents the full set of variables included in the dataset and exploited in our experiments. From many possibilities, we have focused on features collected at admission time, and these are the ones shown in Table 1. This choice is motivated by clinical utility, as patients may present acute worsening of their symptoms without giving hospitals time to prepare in advance. For instance, a health center may reserve oxygen or an extra intensive care unit bed for an additional patient, or even forward the patient to another center. As such, this represents an overall adequate timing to provision resources, as it ensures more time to maneuver available assets.

Table 1. Variables included in our experiments

Variables	
Demographics characteristics	Illegal drug use
Sex at birth	Alcoholism
Age (years)	Current smoker
Comorbidities and lifestyle habits	Ex-smoker
Hypertension	**Clinical characteristics**
Coronary artery disease	Time from symptom onset
Heart failure	Respiratory rate (irpm)
Atrial fibrillation or flutter	Heart rate (bpm)
Stroke	Systolic blood pressure (mm Hg)
Chagas disease	Diastolic blood pressure (mm Hg)
Rheumatic heart disease	Inotrope use
Other cardiovascular disease	Glasgow coma score
No relevant cardiovascular disease	SF ratio
Asthma	FiO2
COPD	**Laboratory**
Pulmonary fibrosis	C reactive protein (mg/L)
Diabetes mellitus	Hemoglobin (g/L)
Obesity (BMI¿30kg/m2)	Leucocytes (109/L)
Cirrhosis	Neutrophils (109/L)
Psychiatry disease	Lymphocytes (109/L)
Chronic kidney disease	Neutrophils-to-lymphocytes ratio
Rheumatologic disease	Platelet count (109/L)
HIV infection	Creatinine (mg/dL)
Cancer	Urea (mg/dL)
Previous organ transplantation	Lactate (mmol/L)
Immunosuppressive condition	Sodium (mmol/L)
Another relevant health condition	Bicarbonate (mEq/L)
Nº comorbidities	pH
Nº cardiovascular comorbidities	pO2 (mmHg)
Nº of comorbidities in different groups	pCO2 (mmHg)

4. Analyzing the Effect of Temporal Drifts on the Brazilian Cohort COVID-19 Dataset

In this section, we aim at characterizing the diverse set of temporal effects that occurred in our COVID-19 hospitalized patients' dataset. We begin by setting up a baseline for the variation in effectiveness on a simple future death prediction problem. For this, we present results of a binary (death vs. non-death) classification task trained on admission variables for two smaller periods (from March 2020 to March 2021, and from March 2021 to November 2021) and the longest possible period (the whole period March 2020 to November 2021). We begin by showing the unexpected drop in effectiveness that is observed when using the full dataset when compared to use specific (shorter) periods of time, then we proceed into analyzing the possible explanations for this fact, especially those related to potential data/concept drifts.

The results of the aforementioned analyses are presented in Table 2. To perform the experiments, we took advantage of the fact that each of our data points consists only of patient admission attributes, ensuring no data leaks occur when performing cross validation. Indeed, the only temporally-related feature in our dataset is the admission time, which is not used for prediction purposes. Accordingly, we proceeded with a 10-fold cross validation procedure while considering the three mentioned views, the two halves of the dataset and the full dataset. The splitting point between the partitions corresponds to roughly the mid point of our dataset and also a point at which vaccination in

Brazil started to reach 2-3% of the population with at least one dose, not being so widespread as to cause an impact.

For each of the 10 test folds in each partition, we measured both micro and macro averages for the F1-score obtained by applying the respective learned model (trained with 8 splits and tuned in one). The F1 score is a metric obtained through the harmonic mean of precision and recall metrics. Micro average combines the F1-score of the 2 classes (death x non-death) without considering class imbalance while the macro average does so by treating all classes as equally important. In the specific case of this learning task, the macro-F1 score is the most appropriate measure due to high class imbalance – approximately 80-20 – which, as shown in Table 4, is present at all time-based partitions. Statistical significance was assessed by performing Wilcoxon's signed-rank test, since we can not assure a normal distribution of the data.

In terms of classification models, we have included: a standard support vector machine (SVM); LASSO and GAM regression, due to their excellent results in (Marcolino et al., 2021); a boosting model (LightGBM), which, according to (Shwartz-Ziv & Armon, 2022) is usually one of the best performing models when applied to tabular data such as our COVID-19 dataset; a deep neural network representative (1D Convolutional neural network), and a Stacking model. Our stacking model is a combination of the output all other models, while using a logistic regression classifier as combination strategy. The hyperparameters tested for each model in the validation split can be found in Table 3. For configuring the stacking parameters, we used a nested cross-validation procedure within the training set, as explained in (Gomes et al., 2021).

Table 2. Cross-validation comparison of different classifiers while learning on 2 years of COVID-19 hospitalized patients' data.

	First Half			
	micro-f1		**macro-f1**	
Stacking	**0.855**	±0,007	**0.739**	±0,018
LightGBM	**0.846**	±0,008	0.723	±0,016
GAM	**0.847**	±0,006	0.720	±0,014
Lasso	0.842	±0,009	0.677	±0,024
SVM	0.839	±0,010	0.691	±0,031
CNN1D	0.815	±0,013	0.693	±0,016
	Second Half			
	micro-f1		**macro-f1**	
Stacking	**0.805**	±0,019	**0.698**	±0,019
LightGBM	**0.807**	±0,018	**0.695**	±0,017
GAM	**0.805**	±0,018	**0.692**	±0,015
Lasso	0.799	±0,020	0.668	±0,018
SVM	**0.804**	±0,021	0.675	±0,022
CNN1D	0.771	±0,019	0.666	±0,018
	Full Dataset			
	micro-f1		**macro-f1**	
Stacking	**0.821**	±0,046	**0.654**	±0,026
LightGBM	**0.825**	±0,043	**0.648**	±0,029
GAM	0.813	±0,051	0.630	±0,019
Lasso	0.809	±0,053	0.595	±0,024
SVM	0.814	±0,046	0.608	±0,024
CNN1D	0.776	±0,050	0.625	±0,019

Table 3. Classification models' hyperparameters

Methods	**Parametrization**
SVM	C : [10-3, 10-2, 10-1, 100, 101, 102]
	Kernel: [linear, rbf, poly, sigmoid]
	class_weight: [None, 'balanced']
Lasso	Alpha: [10-3, 10-2, 10-1, 100, 101, 102]
LightGBM	N-estimators: [10, 50, 100, 200, 500, 1000, 2000]
	learning_rate: [10-3, 10-2, 10-1, 30-1]
	colsample_by_tree: [0.5, 1.0]
CNN	Epochs: Until Early Stop
	learning_rate: $1e-4$
	filter_size: 32
	activation: 'relu'
	num_layers: 4
GAM	No tunning
Stacking	Meta-Classifier: Logistic Regression, Alpha: [102]

As aforementioned, the results from the cross-validation test are shown in Table 2. In this table, bold values indicate results significantly superior within that time partition. Micro and macro f1 scores for 6 classifiers, including the stacking model, are shown, as well as the confidence intervals for these results. Regarding the individual models, overall, the neural network model was the worst performer in all tests, possibly related to the size of our dataset (10K samples for the full data and 5K samples for the halves), which is too small for such huge models. The best individual model was LightGBM, and the best overall model is our combination strategy (Stacking). These two models were statistically tied at micro f1 on all tests, but stacking has a slight edge, being statistically superior in the first period. The high effectiveness of LightGBM in all scenarios is possibly due to the tabular nature of our learning task, as suggested in (). On the other hand, the superior performance of the Stacking model in all scenarios is consistent with recent results (Gomes et al., 2021).

A key observation in this Table is the fact that, for all classifiers, there is a statistically significant drop in effectiveness when using the complete data versus training only in the two halves. For example, the effectiveness of our best model (Stacking), as measured by MacroF1, drops up to 11.5% in the full dataset versus that obtained in the first period. This is true, even though the period which had the best results (the first half) is fully included in the 'complete' dataset. Another interesting observation relates to how the variance (and with it, the size of the confidence intervals) of the results increases as we move from the first to the second half, and then to the full dataset. This higher variability is possibly due to the fact that, in the second half, we have many more interacting factors (such as vaccines and viral variants). Indeed, by considering the full period, we are in fact increasing heterogeneity and the difficulties for the classifier. Although we have more data when combining the two periods (and then, more data to learn from), combining the two heterogeneous partitions resulted in worst results.

Some interesting questions arise. For instance, why do we have important effectiveness variations between the two partitions, and why does combining them yield worse classifiers? As we will show in the following paragraphs, multiple (temporal) factors seem to be implicated in these outcomes.

Table 4. Fraction of patients per class at each time period

	recovered	died
Until march-2021	79.85%	20.15%
Until november-2021	78.28%	21.72%
March-november-2021	75.34%	24.66%

Among the potential variations that could lead to such effectiveness degradation, one of the most directly observable is **class distribution** drift. The COVID-19 pandemic induced multiple potential class drifts, such as drift in mortality and hospitalization rates. In the specific case of our dataset, we measured mortality over time. These results are shown in Figure 2. The Figure shows the monthly moving average of mortality for the past month at each day in our dataset. On the X-axis, we show time, as measured by means of patient admission dates, while on the Y-axis, we show mortality rates over the past 30 days. As we can see, between the extremes of this curve there is a maximum of over 30% variation in mortality, going from just under 17% up to 23%, over the course of a few months.

Though simple to understand from a clinical perspective, the case fatality rate drift has important effects in the learning process, since classifiers will tend to learn to reproduce predictions somewhat close to the training distributions, which are far from constant over time. Additionally, these rather large drifts of more than 30% occur in the minority class, which can cause even more confusion for the predictor to be learned. This is mainly true if the learning algorithm explores some information from prior distributions, either directly such as Naive Bayes, or indirectly, such as LGBM that exploits some type of sampling from the training data.

Additional sources of temporal drifts in the data include **feature drifts** which can be both semantic and distributional. These drifts, on their turn, may result in either gains or losses of correlation between the input and the target variables. To test for this kind of temporal effect, we further split the data into trimesters and measured the Pearson correlations between the overall top-5 most predictive admission features for future lethality and the death outcome in each trimester. We opt for this additional split in trimesters to better analyze how these effects change over time, as opposed to just comparing the correlations in the two halves and in the full dataset.

The Pearson correlation metric shown in Figure 3 is the ratio between the covariance of two variables and the product of their standard deviations. As such, it is essentially a normalized measure of their covariance, yielding results

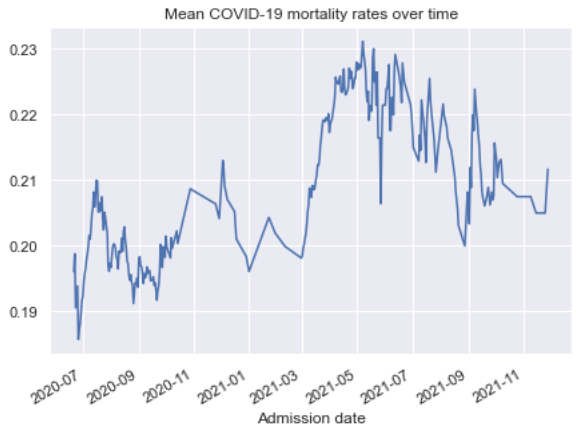

Figure 2. Proportion of hospitalized patient deaths over time.

between -1 and +1 that represent how much linearly correlated the two variables are. The formula for this metric is shown in Equation 1. The intuition behind this analysis lies in attempting to capture how much variables correlate to the target and how their relative ranking changes over time. Significant variations observed in this analysis, particularly in how correlated variables are and in their relative ranking, imply changes in the typical profile of the target classes when compared to the overall distribution.

$$\rho(X,Y) = \frac{cov(X,Y)}{\sigma_X \sigma_Y} \tag{1}$$

Equation 1: Pearson Correlation

In the Pearson correlation analysis of Figure 3, we can see that both the relative ranking and the absolute correlation of each feature changes over time. On the X-axis, we show trimesters of observations, while on the Y-axis, we show absolute Pearson correlation scores. Notice, for instance, how at the earlier stages of the pandemic, Age was the single best predictor of future death out of the 5 best variables, while on the last one, it was the worst. Possibly, this effect was the result of an interaction between elder patients having a naturally higher risk of death, resulting in age being the best predictor in the first trimester. Elderly patients were also the first vaccinated, leading to non-vaccinated elderly patients dying more often than their vaccinated pairs and acting as noise with respect to patient's age as a feature for death prediction.

This sort of effect may induce wrong future predictions, as a classifier may learn that, for instance, patient age is highly predictive of future death, but, as previously stated, elderly patients received vaccines first, resulting in them having an overall smaller risk when compared to relatively younger peers and other non-vaccinated groups.

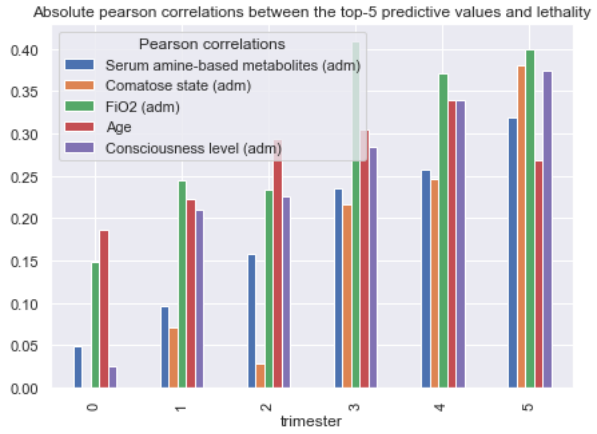

Figure 3. Absolute Pearson correlations over time for the 5 overall most correlated features in our dataset. Notice that the relative rankings and the most predictive features change over time changes.

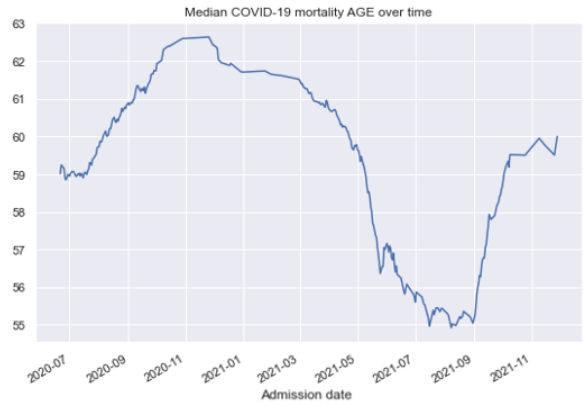

Figure 4. Median age of COVID-19 dying patients over time.

It is interesting to observe the particular case of patient age as a predictive feature for future COVID-19 related deaths, as it was the most predictive in the early periods (e.g., trimester 0 and 2) and became less discriminative in later periods (e.g., trimester 5). In our experiments, we measured the median age of death for patients that died from COVID-19 and show the results on Figure 4. In this Figure, we can see that at the earlier stages of the pandemic, older patients (median age between 60-63 years) were the typical profile for deaths. As time went by, vaccines were developed and older patients were prioritized for vaccination in Brazil. This resulted in a subsequent fall of median death age, which achieved a lowest value of around 55 years of age by September,2019. COVID-19 vaccines had a known effect of not necessarily preventing infections but actually reducing casualty rates. However, as time advances and vaccination progressed, new variants continued to emerge. This culminated in a subsequent rise of median mortality ages, as elder patients were (once again) susceptible to some of these new variants.

To further analyze how the profile of a typical dying COVID-19 patient changes over time, we calculated the mean feature values over all dying patients (i.e. their centroid) in our dataset, and assessed the mean similarity to that centroid over time. The general profile of a dying patient is represented by a 45-position vector where each position contains attributes measured at hospital admission time. The general centroid has the mean value for each feature for all dying patients. This procedure is described in Equation 2. From this, we analyzed the mean cosine similarity between this general centroid and a similar vector representing patients who have died at each particular point in time.

The results from this analysis are shown in Figure 6. From the Figure, we can see that there is a trend towards deviation

$$centroid = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{2}$$

Equation 2: Dying patient's general centroid

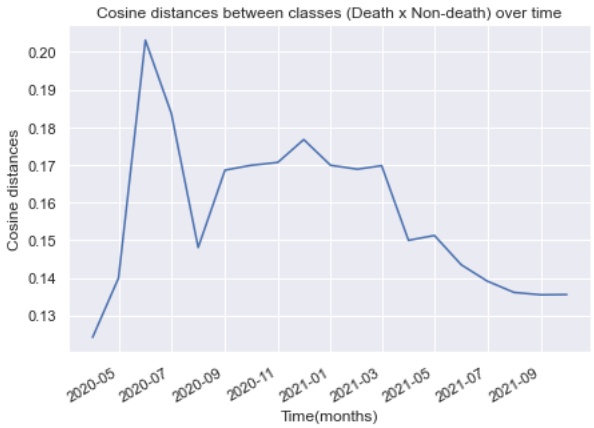

Figure 5. Mean dissimilarity between classes (Death x Non-death) over time.

and approximation from this centroid over time, with the highest dissimilarities occurring at the earlier stages of the pandemic, and higher similarities occurring around July-2021. This pattern shows that the 'dying patients' profile' changes over time, and a well tuned classifier trained on just one period will most certainly underperform when applied at another period. Notice how the "first big shift" occurs right at the onset of the pandemic, around May-July 2020, when some states began experiencing exponential growth in cases, causing further deaths of less severely ill patients. The second major profile change occurs around March-2021, when vaccination starts to gain traction in the country, which initially shifted the dying patient's profile towards the younger, unvaccinated patients.

Another important factor that affects how classification tasks behave over time is the **evolution of the relationships among classes**, as in how similar or dissimilar they are over time. We analyze this particular trait on Figure 5. In the Figure, the X-axis shows time (in months), while the y-axis shows the mean cosine distance between patients in both classes. To generate this plot, we move a sliding window that captures each month in our database, and filter, in that view, the patients that died and the ones that did not die. In this Figure, we can see that patients that died were also more dissimilar at the earlier stages of the pandemic, which partly explains the better results in the first half of our dataset when compared to the second half and to the full dataset. As time passes younger patients started to die from COVID-19, from multiple reasons, from being unvaccinated to being more exposed to the disease, their characteristics became more similar to those of the 'recovered' group.

An interesting exercise that relates to the results in Figure 5 is that of characterizing why the first partition yields more effective classifiers than the second one. Going back to Table 2, we can see that the first partition yelds the best results, followed by the second one, while both yield better results than the full dataset (looking at macro-f1). A major reason for this, from a temporal drift standing point, is simply the fact that classes are more dissimilar on the first partition. The dying patients on the second partition are younger and more similar to the ones that do recover, versus relatively older and more diseased patients in the first partition.

To conclude, we have shown that the task of predicting future COVID-19 mortality upon hospitalization has suffered multiple temporal drifts over time, including class distribution shifts, feature importance changes and even the modification of the overall risk profile for death. Such effects impair classification effectiveness, and the solution to these problems is not trivial, as there is a trade-off between selecting less samples form a more recent time to re-train a classifier – and hence having less information to learn from – and selecting more samples to learn from – but incurring at the danger of incorporating noise into the model due to the multiple temporal shifts in the data.

5. Conclusion

In this work, we presented evidence of how several types of temporal effects have impacted important characteristics of the COVID-19 pandemic, including the dying patient's profile. For instance, we have shown that mortality rates have changed over time, possibly due to the interaction between multiple vaccines being applied targeting specific patterns (e.g., specific age groups and populations first). Vaccines seem to have an important influence on some of the temporal drifts we observed in this pandemic. We have also demonstrated the classification effectiveness degradation

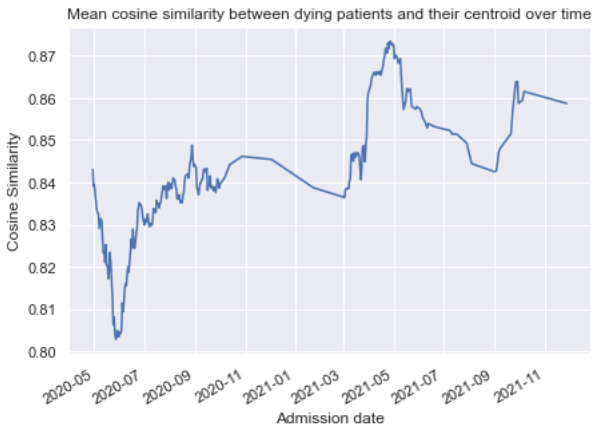

Figure 6. Mean cosine similarity between dying patients and their overall centroid over time.

that emerges out of using data containing all of these heterogeneous patterns and shown how the correlation between different predictors and the target variable (in our case, future death) can change over time. In order to properly address these temporal effects, we may employ and adapt some strategies already proposed in the literature, such as sample selection (Rocha et al., 2008) and sample weighting (Salles et al., 2010). These strategies can minimize the temporal effects by ensuring that the most relevant samples will have the highest relative impact on the learning algorithm. These strategies also aim at maximizing effectiveness under the trade-off between having less data to learn from and learning from the most relevant (and possibly most recent) samples.

In the future, we will explore the impact of both sample selection and sample weighting strategies when applied to COVID-19 data aiming at mitigating the temporal drifts that impact machine learning models applied in this context. We also intend to work on meta-features derived from the population in a given period of time to minimize the temporal effects. Finally, we intend to test these solutions on other COVID-19 datasets besides the one described in this work.

Acknowledgments

This work is supported by CAPES, CNPq (grant number 465518/2014-1) and Fapemig (grant number APQ-01154-21).

References

Carbonell, R., Urgelés, S., Rodríguez, A., Bodí, M., Martín-Loeches, I., Solé-Violán, J., Díaz, E., Gómez, J., Trefler, S., Vallverdú, M., et al. Mortality comparison between the first and second/third waves among 3,795 critical covid-19 patients with pneumonia admitted to the icu: A multicentre retrospective cohort study. *The Lancet*

Regional Health-Europe, 11:100243, 2021.

Gama, J. and Zhang, G. Learning under concept drift: A review. *IEEE TKDE*, 31(12), 2019.

Gomes, C., Gonçalves, M. A., Rocha, L., and Canuto, S. D. On the cost-effectiveness of stacking of neural and non-neural methods for text classification: Scenarios and performance prediction. In *Association for Computational Linguistics: ACL/IJCNLP 2021*, pp. 4003–4014, 2021.

Haque, A., Khan, L., Baron, M., Thuraisingham, B., and Aggarwal, C. Efficient handling of concept drift and concept evolution over stream data. In *ICDE*, pp. 481–492, 2016.

Iftimie, S., López-Azcona, A. F., Vallverdú, I., Hernández-Flix, S., De Febrer, G., Parra, S., Hernández-Aguilera, A., Riu, F., Joven, J., Andreychuk, N., et al. First and second waves of coronavirus disease-19: A comparative study in hospitalized patients in reus, spain. *PloS one*, 16(3): e0248029, 2021.

Ikemura, K., Bellin, E., Yagi, Y., Billett, H., Saada, M., Simone, K., Stahl, L., Szymanski, J., Goldstein, D., and Gil, M. R. Using automated machine learning to predict the mortality of patients with covid-19: Prediction model development study. *JMIR*, 23(2):e23458, 2021.

Jassat, W., Mudara, C., Ozougwu, L., Tempia, S., Blumberg, L., Davies, M.-A., Pillay, Y., Carter, T., Morewane, R., Wolmarans, M., et al. Difference in mortality among individuals admitted to hospital with covid-19 during the first and second waves in south africa: a cohort study. *The Lancet Global Health*, 9(9):e1216–e1225, 2021.

Jung, C., Excoffier, J.-B., Raphaël-Rousseau, M., Salaün-Penquer, N., Ortala, M., and Chouaid, C. Evolution of hospitalized patient characteristics through the first three covid-19 waves in paris area using machine learning analysis. *Plos one*, 17(2):e0263266, 2022.

Klinkenberg, R. Learning drifting concepts: Example selection vs. example weighting. *Int. D. Analysis*, 8(3): 281–300, 2004.

Kolter, J. Z. and Maloof, M. A. Dynamic weighted majority: An ensemble method for drifting concepts. *JMLR*, 8: 2755–2790, 2007.

Marcolino, M. S., Pires, M. C., Ramos, L. E. F., Silva, R. T., Oliveira, L. M., Carvalho, R. L., Mourato, R. L. S., Sánchez-Montalvá, A., Raventós, B., Anschau, F., et al. Abc2-sph risk score for in-hospital mortality in covid-19 patients: development, external validation and comparison with other available scores. *Int. J. of Infectious Diseases*, 110:281–308, 2021.

Moreno-Torres, J. G., Raeder, T., Alaiz-Rodríguez, R., Chawla, N. V., and Herrera, F. A unifying view on dataset shift in classification. *Pat. Recog.*, 45(1):521–530, 2012.

Mourao, F., Rocha, L., Araújo, R., Couto, T., Gonçalves, M., and Meira Jr, W. Understanding temporal aspects in document classification. In *WSDM ´08*, pp. 159–170, 2008.

Paiva, B. B. M., Pereira, P. D., Andrade, C. M. V., Gomes, V. M. R., Gonçalves, M. A., and Marcolino, M. S. Potential and limitations of machine meta-learning (ensemble) methods for predicting covid-19 mortality in a large in hospital brazilian dataset. *Under evaluation*, 2022.

Rocha, L., Mourão, F., Pereira, A., Gonçalves, M. A., and Meira Jr, W. Exploiting temporal contexts in text classification. In *CIKM ´08*, pp. 243–252, 2008.

Salles, T., Rocha, L., Pappa, G. L., Mourão, F., Meira Jr, W., and Gonçalves, M. Temporally-aware algorithms for document classification. In *SIGIR ´10*, pp. 307–314, 2010.

Salles, T., Rocha, L., Gonçalves, M. A., Almeida, J. M., Mourão, F., Meira Jr, W., and Viegas, F. A quantitative analysis of the temporal effects on automatic text classification. *JASIST*, 67(7):1639–1667, 2016.

Salles, T., Rocha, L., Mourão, F., Gonçalves, M., Viegas, F., and Meira Jr, W. A two-stage machine learning approach for temporally-robust text classification. *Inf. Systems*, 69: 40–58, 2017.

Shwartz-Ziv, R. and Armon, A. Tabular data: Deep learning is not all you need. *Information Fusion*, 81:84–90, 2022.

Telenti, A., Arvin, A., Corey, L., Corti, D., Diamond, M. S., García-Sastre, A., Garry, R. F., Holmes, E. C., Pang, P. S., and Virgin, H. W. After the pandemic: perspectives on the future trajectory of covid-19. *Nature*, 596(7873): 495–504, 2021.

Zeiser, F. A., Donida, B., da Costa, C. A., de Oliveira Ramos, G., Scherer, J. N., Barcellos, N. T., Alegretti, A. P., Ikeda, M. L. R., Müller, A. P. W. C., Bohn, H. C., et al. First and second covid-19 waves in brazil: A cross-sectional study of patients' characteristics related to hospitalization and in-hospital mortality. *The Lancet Regional Health-Americas*, 6:100107, 2022.