medRxiv preprint doi: https://doi.org/10.1101/2020.04.17.20069666; this version posted January 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Predicting the Epidemic Curve of the Coronavirus (SARS-CoV-2) Disease (COVID-19) Using Artificial Intelligence

László Róbert Kolozsvári[1]*, Tamás Bérczes[2], András Hajdu[2], Rudolf Gesztelyi[3], Attila Tiba[2], Imre Varga[2], Ala'a B. Al-Tammemi [1,4], Gergő József Szőllősi[1], Szilvia Harsányi[4,5], Szabolcs Garbóczy[4,6], Judit Zsuga[5]

[1]Department of Family and Occupational Medicine, Faculty of Medicine, University of Debrecen, Debrecen, Hungary

[2]Faculty of Informatics, University of Debrecen, Debrecen, Hungary.

[3]Department of Pharmacology and Pharmacotherapy, Faculty of Medicine, University of Debrecen, Debrecen, Hungary

[4]Doctoral School of Health Sciences, University of Debrecen, Debrecen, Hungary

[5]Department of Health Systems Management and Quality Management in Health Care, Faculty of Public Health, University of Debrecen, Debrecen, Hungary

[6]Department of Psychiatry, University of Debrecen, Debrecen, Hungary

*Correspondence: kolozsvari.laszlo@med.unideb.hu; Tel.: +36 70 384 5265

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2020.04.17.20069666; this version posted January 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

ABSTRACT: The aim of our study was to predict the epidemic curves (daily new cases) of COVID-19 pandemic using Artificial Intelligence (AI)-based Recurrent Neural Networks (RNNs), then to compare and validate the predicted models with the observed data. We used the publicly available datasets from the World Health Organization and Johns Hopkins University to create a training dataset, then we used RNNs with gated recurring units (Long Short-Term Memory) to create two prediction models. Information collected in the first t time-steps were aggregated with a fully connected (dense) neural network layer and a consequent regression output layer to determine the next predicted value. We also used Root Mean Squared Logarithmic Errors (RMSLE) to compare the predicted models with the observed data. The result of our study underscores that the COVID-19 pandemic is a propagated source epidemic, therefore repeated peaks on the epidemic curve are to be anticipated. Besides, the errors between the predicted and validated data and trends seems to be low. The influence of this pandemic is significant worldwide and has already impacted our daily life. Decision makers must be aware, that even if strict public health measures are executed and sustained, future peaks of infections are possible.

Keywords: COVID-19; SARS-CoV-2; Artificial Intelligence; Epidemic Curve; Prediction; Recurrent Neural Networks.

medRxiv preprint doi: https://doi.org/10.1101/2020.04.17.20069666; this version posted January 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

1.INTRODUCTION

1.1 Coronavirus

High and low pathogenic species may be distinguished within the coronavirus family, with the former including 4 viruses that are responsible for 10-30% of mild upper respiratory diseases (e.g. common cold), and the latter known to cause a more severe form of acute lung injury: SARS-CoV (Severe Acute Respiratory Syndrome) and MERS-CoV (Middle East Respiratory Syndrome) [1]. SARS-CoV originated in Guangdong Province, China, and started to spread in 2002, causing over 8,000 cases in 29 countries around the world, with a crude fatality rate of 10% [2–4]. The virus has spread to Hong Kong in 2003 causing an outbreak of severe acute respiratory syndrome (SARS). A novel coronavirus was isolated and was suggested to be the primary cause of the infections [5]. A few years later, in 2007 Cheng et. al issued a warning that "the presence of a large reservoir of SARS-CoV-like viruses in horseshoe bats, together with the culture of eating exotic mammals in southern China, is a time bomb" [4]. MERS-CoV began spreading in Saudi Arabia in 2012 and, to the time of writing this, has led to a total of 2519 laboratory-confirmed cases in several countries around the world [6,7]. Its case-fatality rate reached 37.1% over the course of the past 8 years [7].

1.2 Coronavirus Disease – 2019 (COVID-19)

The current form of the Severe Acute Respiratory Syndrome noted as COVID-19, is caused by a new variant of formerly known highly pathogenic *Coronaviridae*. The infection allegedly began to spread from a market in Wuhan, the capital of Hubei province, China, at the end of 2019 [8,9]. Early genome sequencing has found that the new virus, which was named SARS-CoV-2 by the International Committee on Taxonomy of Viruses, showed a 79.6% homology with SARS-CoV, and has 96% sequence identity with bat coronavirus suggesting a common origin from SARSr-CoV (severe acute respiratory syndrome-related coronavirus). Also, the suspected host was thought to be a bat species, *Rhinolophus affinis* (a horseshoe bat), but the SARS-COV-2 probably needs an intermediate host [9,10].

medRxiv preprint doi: https://doi.org/10.1101/2020.04.17.20069666; this version posted January 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Symptoms associated with the COVID-19 include fever (83%), cough (82%), shortness of breath (31%), muscle aches (11%), confusion (9%), headache (8%), sore throat (5%), runny nose, chest pain, diarrhea, nausea and vomiting [11]. According to a meta-analysis that compiled data from more than 50 000 patients, the incidence of fever (0.891, 95% confidence interval (CI): [0.818; 0.945]) and cough (incidence of 0.722, 95% CI: [0.657; 0.782]) were the highest, respectively, followed by muscle soreness and fatigue [12]. The incubation period of the virus was estimated to be between 1-14 days (5 days on average) [13]. There is no definite data concerning the transmissibility of the virus. However, several transmission routes have been identified including respiratory droplets/aerosols, direct contact with virally-contaminated objects, and possibly fecal-oral transmission [14].

It seems probable that those with a fulminant disease are most infectious, but reports have identified asymptomatic and pre-symptomatic virus shedding as well. There is also a lack of definite data regarding tertiary and quaternary spreading among humans, but it seems probable that the person who has been exposed to the infection has acquired some (at least temporary) immunity [15]. According to data from the World Health Organization (WHO), there were 91,816,091 confirmed cases and 1,986,871 fatalities globally as of January 15, 2021, which corresponds to a case-fatality rate of 2.164% [16].

R0, the basic reproduction number, denoting the transmissibility of a virus that indicates the average number of new infections induced by an infectious person in a susceptible, infection naïve population. The transmissibility of the virus was apparently underestimated initially by the WHO with R0 suggested to range between 1.4 and 2.5. More recent reports have indicated higher R0 values around 3 (with the mean and median R0 for published estimates being 3.28 and 2.79, respectively) [13,17].

medRxiv preprint doi: https://doi.org/10.1101/2020.04.17.20069666; this version posted January 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

1.3 The Daily Number of Newly Diagnosed Infections - Epidemic Curves

The initial epidemic curves of the COVID-19 outbreak from Hubei, China showed a mixed pattern, indicating that early cases were likely from a continuous common source e.g., from several zoonotic events in Wuhan, followed by secondary and tertiary transmission providing a propagated source for the later cases [18].

The propagated (or progressive source) epidemic curve visualizes the spread of an infectious agent that may be transmitted from human to human starting from with a single index case, that continues to further infect other individuals. This shows up as a series of peaks on the epidemic curve, that starts with the index case, followed by successive waves of the infection set apart with respect to the incubation period of the pathogen. The waves continue to follow each other, until appropriate mitigation measures, prevention, or treatment are implemented, or the pool of the susceptible population becomes infected. This is a theoretic curve, that is generally influenced by lots of other factors [18].

Several studies investigated the impact of different interventions with respect to minimizing contact rates in the population in order to slow the infection spread, minimize COVID-19 mortality rates and health care utilization, or to suppress the epidemic per se. Flattening the curve by reducing peak incidence may limit overall case fatality rates. Nevertheless, most of the forecasts and simulations thus far started out from Bell-shaped curves, that fail to account for the progressive nature of the current outbreak given the known secondary, tertiary even quaternary transmissibility of the virus. Taking this into account it is suggested that the number of cases will rise once again after pandemic control measures are no longer in effect [19].

1.4 Prediction

There are different mathematical models that may demonstrate and predict the dynamics of different infectious diseases [20]. These models, used to simulate the dynamics of infectious diseases, may be based on statistical, mathematical, empirical, or machine learning methods [21].

medRxiv preprint doi: https://doi.org/10.1101/2020.04.17.20069666; this version posted January 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

The first attempts to use Artificial Intelligence (AI) in medicine were made in the 1970s. Initially, AI was used to implement programs to help clinical decision-making, but to date, its use is gaining more and more widespread acceptance in biomedical sciences [22]. One class of AI, a form of artificial neural networks, the *Recurrent Neural Networks (RNNs) with Long Short-Term Memory (LSTM)* were previously used to model and forecast the influenza epidemic, with strong competitiveness and reliable results [23–25].

1.5 Aim of the Study

The current study aimed to use the publicly available official COVID-19 data as a training dataset, to predict the possible outcomes of the COVID-19 pandemic (epidemic curve of new cases) using AI-based RNNs, and further, to compare the predictions with the observed data.

2. MATERIALS AND METHODS

2.1 Data

We used the publicly available datasets from the WHO and Johns Hopkins University for the following countries to create the ***training dataset***: Austria, Belgium, China (Hubei) , Czech Republic, France, Germany, Hungary, Iran, Italy, Netherlands, Norway, Portugal, Slovenia, Spain, Switzerland, United Kingdom (UK) and the United States of America (USA) [15,26]. Given that most infected people in China were from Hubei province, only data from that province was included. For each country, the date of the first reported infection was set as ***Day 1*** for the disease time scale. (**Figure 1**).

medRxiv preprint doi: https://doi.org/10.1101/2020.04.17.20069666; this version posted January 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

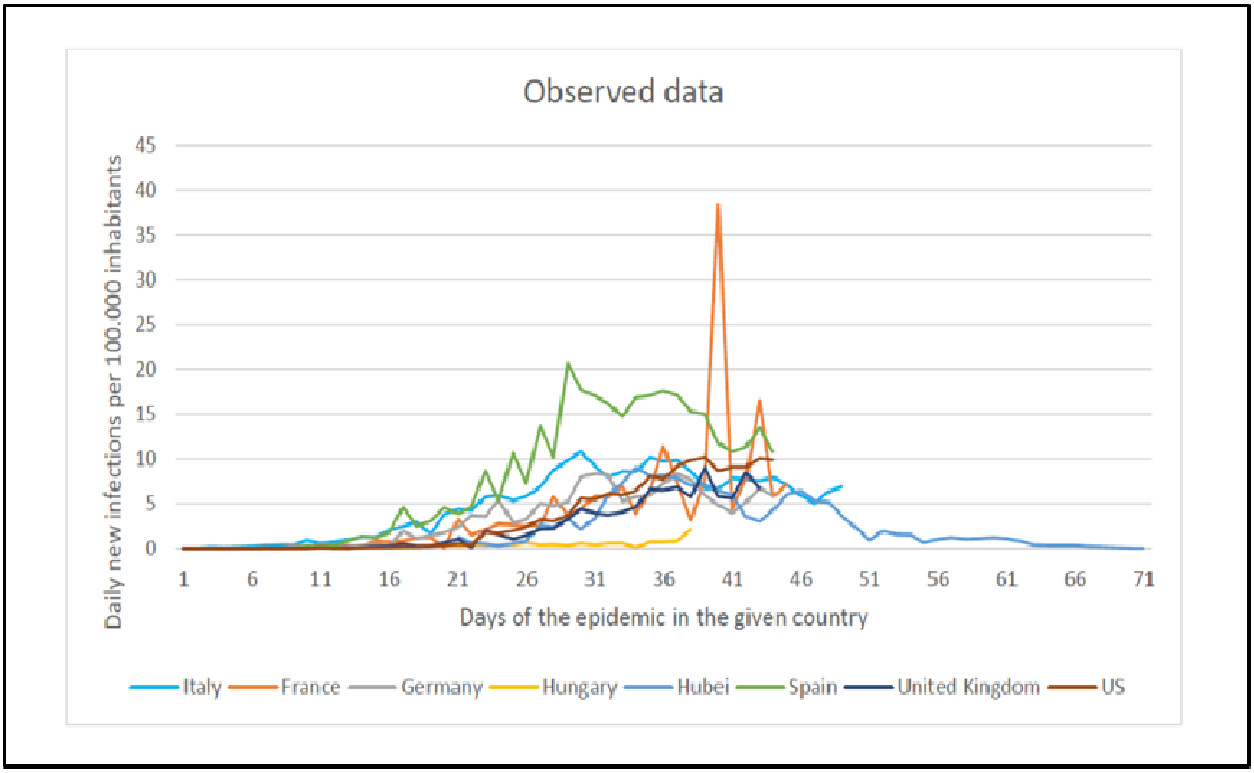

Figure 1. The Historical Datasets of Different Countries

medRxiv preprint doi: https://doi.org/10.1101/2020.04.17.20069666; this version posted January 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

When determining the date of the first illness, point-source outbreaks were omitted (e.g., those cases where single verified cases were isolated, and no further transmission has occurred). This was important to avoid distortion of the propagated epidemic curves. In Belgium, for example, the first illness occurred on February 02, 2020 and there were no further cases reported for up to 26 days. The next illness occurred on March 01, 2020. Inclusion of the early case from February would contribute to a false learning rule for the AI, hence corrupting the results. As for Hubei Province, the first officially available data is of January 22, 2020. This cannot be considered as the first day of the illness, thus the first infection was arbitrarily defined to occur on January 01, 2020. To account for the extreme variability of daily incident cases reported which probably reflects delays in reporting procedures, a moving average was used (covering 3 days) for the Hubei dataset.

Accordingly, an epidemic curve was obtained for each country with a time series where the first day denotes the day of the first confirmed case, and each successive day indicating the number of newly confirmed cases that day. To account for the country-specific differences in the size of the population, the number of daily new cases was normalized for 100.000 inhabitants in each country. The observation period varies for each country, given the difference of time elapsed since the disease initiation in that country. Accordingly, the longest time series covers the observation period of 90 days. e.g., in Hubei, with the first 22 days lacking valid data and the next 68 days having data. The shortest observation period was in Slovenia with only 30 days.

The training data set was obtained by averaging the daily incidence rates per 100 000 inhabitants across the 17 countries included, for each day in the time series. When calculating the average, missing data was left blank, i.e., NULL, e.g., countries that did not contain data for a specific day were excluded from the calculation of average. The resulting ***training data*** set is shown in **Figure 1**. It should be noted that the first part of the data set (up to the initial 30 days since **Day 1** of the epidemic) contains data for almost all the countries listed, whereas the end of the data set contains only data from Hubei. (**Figure 2**)

medRxiv preprint doi: https://doi.org/10.1101/2020.04.17.20069666; this version posted January 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

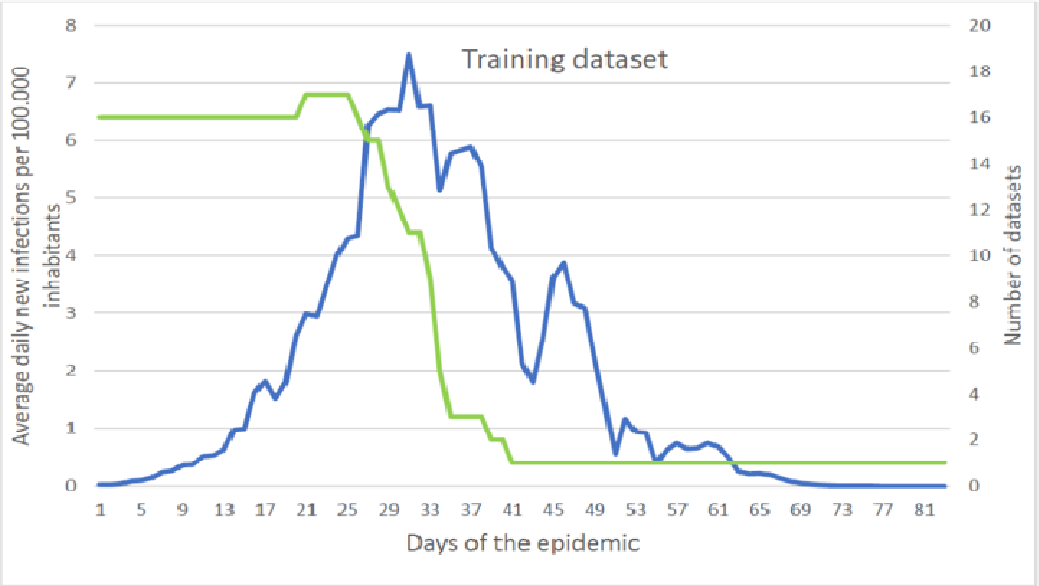

Figure 2. The training dataset. Average daily new infections per 100.000 inhabitants (line) and the Number of datasets (line)

2.2 RNNs-Based Model for Prediction

The state-of-the-art for time series analysis is AI-based analytic tools, which have the best prediction performance. Recurrent Neural Networks (RNNs) are specifically designed to cope with sequential input, characteristic of textual or temporal data [24]. This architecture is a neural network-based architecture, that contains hidden layers chained according to the time step, with a possibility to predict the next sequence element(s). A time series has a special temporal form, where the input to the i-th hidden layer is at the i-th time-step that has a corresponding x(i) observation. In its original form, a simple RNN tries to predict the next sequence element, however, for the purposes of the current analysis, an encoder-decoder variant is a more natural choice, similarly to machine translation [27].

medRxiv preprint doi: https://doi.org/10.1101/2020.04.17.20069666; this version posted January 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

For our specific scenario this means that during the encoder phase including time steps 1,…,t the RNN is fed with the already known time series data (the average of the number of new cases normalized to 100 000 inhabitants for day 1…t, respectively), followed by prediction in the decoder phase for the future time steps t+1,…,T. In our analysis, T=t+1=90 days is the longest known (Hubei) time interval. Since this covers quite a long data sequence, we have used gated recurring units (namely Long Short-Term Memory – LSTM units) in compliance with the general recommendations [25]. **Figure 3** depicts our RNN architecture showing how unknown time series elements are predicted. **Figure 3** also shows how the information collected in the first t time-steps are aggregated with a fully connected (dense) neural network layer and a consequent regression output layer to determine a predicted number of new patients as x(t+1).

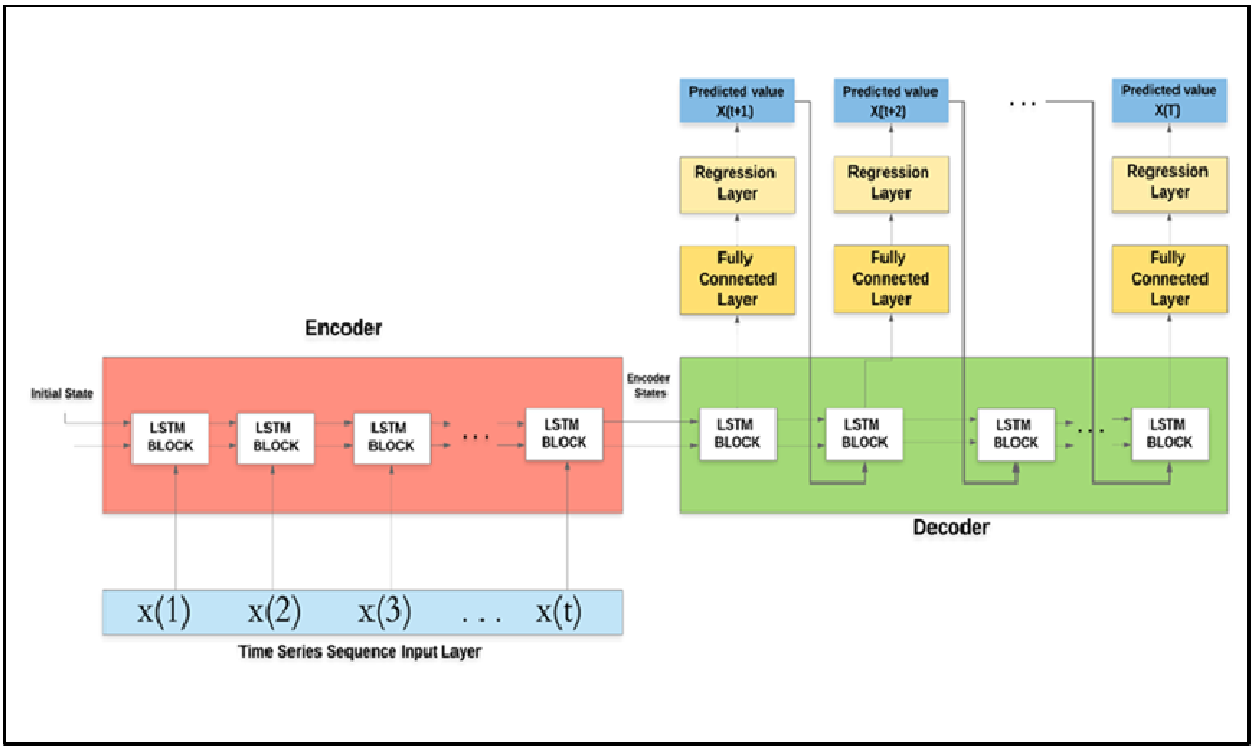

Figure 3. The Recurrent Neural Network (RNN) Architecture used for Prediction.

medRxiv preprint doi: https://doi.org/10.1101/2020.04.17.20069666; this version posted January 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

The training data was described in the previous section. To assess possible specificities regarding the countries two approaches were used for prediction:

•**Prediction 1**: An algorithm to update the training step and subsequent prediction was formulated. This update step is based on the general recommendations of transfer learning that considers the already known time interval for the given country and re-training is done in small increments of the RNN network accordingly [28]. Thus, we start predicting the first unknown element x(t+1) from the last 5% of the known data, and the same principle is applied to each subsequent element. Moreover, after each prediction step our RNN architecture was re-trained and the subsequent elements were predicted with this updated RNN.

•**Prediction 2**: We start predicting the first unknown element x(t+1) from the last known x(t), and all the subsequent elements are predicted only from the preceding ones. Here the rules depicted from the training data set are used, not retraining occurs.

The intuitive interpretations of the difference between Prediction 1 and Prediction 2 are as follows. Prediction 2 makes its predictions utilizing the information derived from the training data set, reflective of the trends in the average time series. It follows that predictions will comply primarily with the Hubei time series, especially in the far future. Therefore Prediction 2 shows the highest fidelity to the country-specific future scenario if the approach to mitigate the epidemic is similar to that in Hubei. Accordingly, this scenario is also reflective of a country-specific future state given the practices of Hubei were followed in said country. On the other hand, Prediction 1 is yielded after the neural network is retrained after any prediction, providing more valid insight into what is expected if the country goes on with the mitigation practices seen during the observation period. The architecture was trained in 250 epochs with a total number of 100 hidden LSTM layers, to prepare a bit for prediction also after T=90 days. Naturally, the length of the RNN can be freely increased later.

medRxiv preprint doi: https://doi.org/10.1101/2020.04.17.20069666; this version posted January 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

2.3 Validation

For the learning dataset, we used the data from the first pandemic wave. That is, we took the available factual data from the first case reported in a country until April 10, 2020. Based on that, we have made the above mentioned two predictions (1 and 2). Moreover, for the validation process, we used the factual data of the first wave. By country, we considered 85-90 days from the first case reported. Thus, the number of days predicted varied from country to country in the same way as for the learning dataset. The amount of Root Mean Squared Logarithmic Errors (RMSLE) was used for validation. In our analysis, the possible bias regarding the different ratios between the observed and predicted values are interpreted using the RMSLE. Let **n** be the number of days used for validation. Let **p_1i** and **p_2i** be the number of new cases per day obtained using the two prediction methods in the examined time interval and let **a_i** be the actual data for the given days. **Err1** and **Err2** will be RMSLE for Prediction 1 and Prediction 2, respectively, where:

$$Err_1 = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log(p_{1i}+1) - \log(a_i+1))^2}$$

$$Err_2 = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log(p_{2i}+1) - \log(a_i+1))^2}$$

medRxiv preprint doi: https://doi.org/10.1101/2020.04.17.20069666; this version posted January 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

3. RESULTS

This section shows the outcomes for **Prediction 1** and **Prediction 2** for the individual country-level data for France, Germany, Hungary, Italy, Spain, USA, and UK (**Figures 4-10**). For each main graph, the small graph in the upper right corner contains the daily error values calculated for the predictions. The more accurate the prediction, the smaller the RMSLE error. It should be noted that if the error function is parallel to the x-axis, it means that the trend of the prediction is the same as the real trend, only at a lower or higher scale. Also, the total errors for the entire investigated period, the summarized mean of the predictions (RMSLE) by country, is shown in **Table 1**.

medRxiv preprint doi: https://doi.org/10.1101/2020.04.17.20069666; this version posted January 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

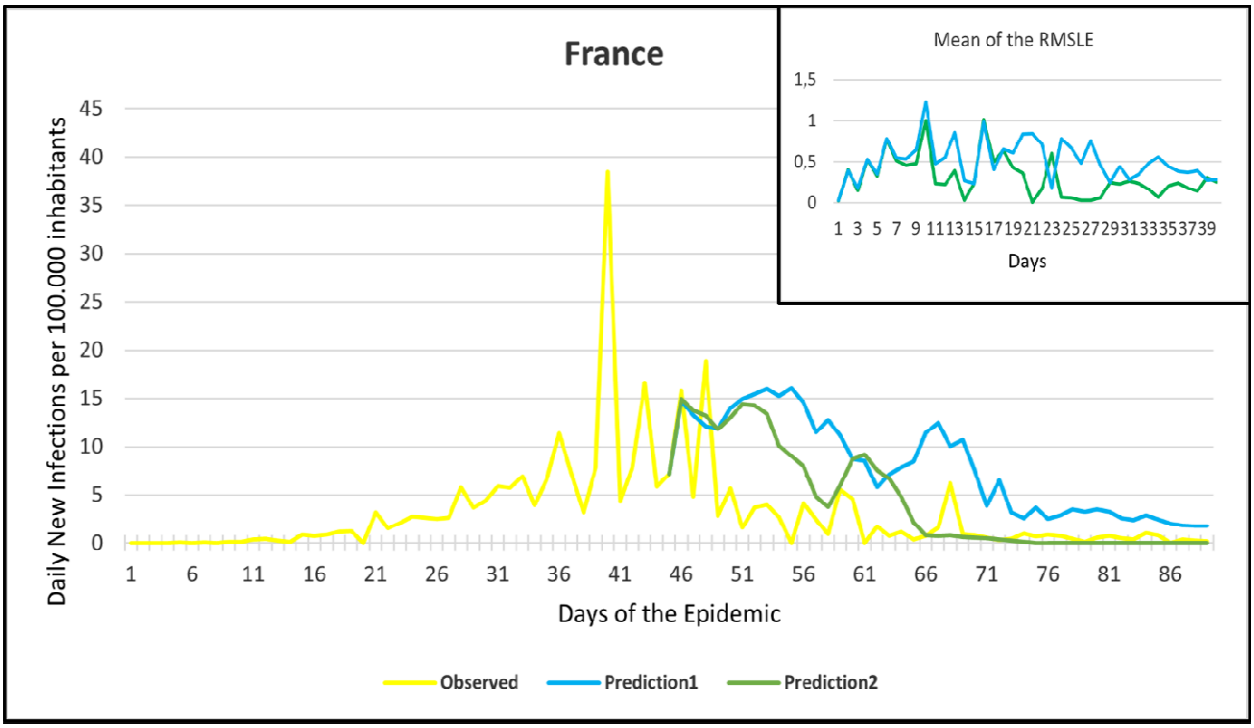

Figure 4. Observation and predictions for France. The small graph in the upper right corner shows the daily error values calculated for the predictions.

medRxiv preprint doi: https://doi.org/10.1101/2020.04.17.20069666; this version posted January 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

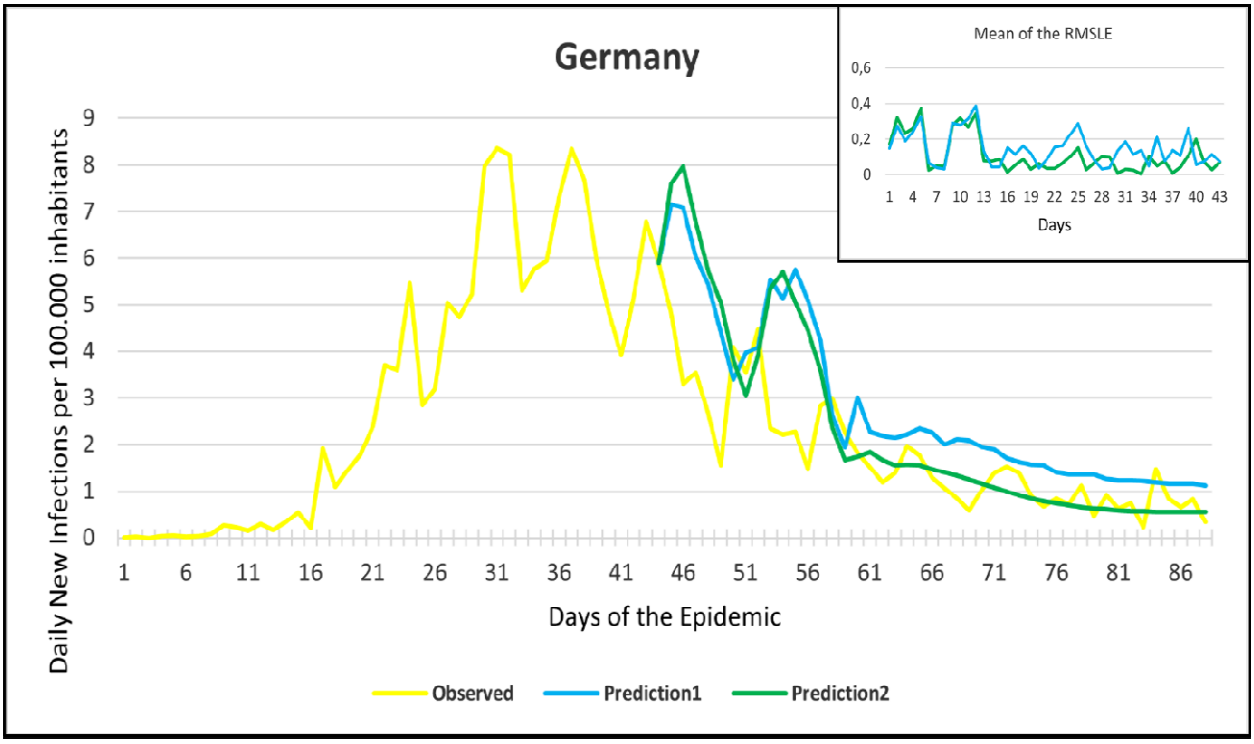

Figure 5. Observation and predictions for the Germany. The small graph in the upper right corner shows the daily error values calculated for the predictions.

medRxiv preprint doi: https://doi.org/10.1101/2020.04.17.20069666; this version posted January 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

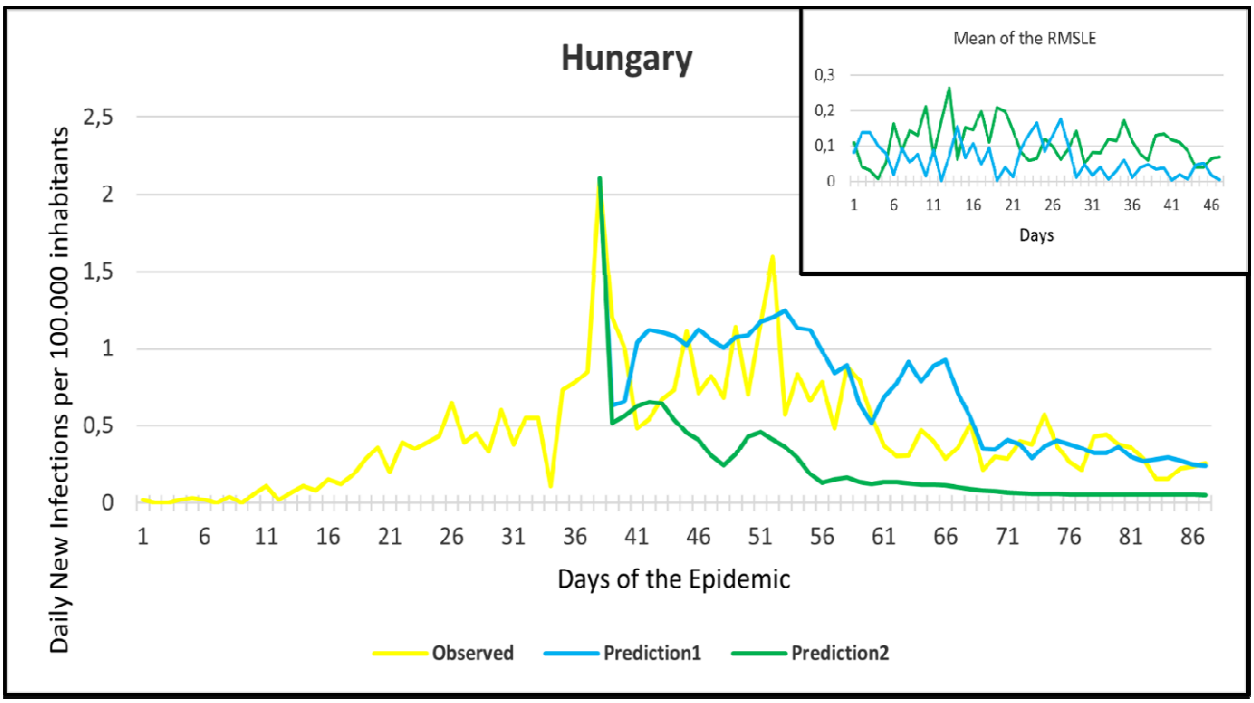

Figure 6. Observation and predictions for Hungary. The small graph in the upper right corner shows the daily error values calculated for the predictions.

medRxiv preprint doi: https://doi.org/10.1101/2020.04.17.20069666; this version posted January 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

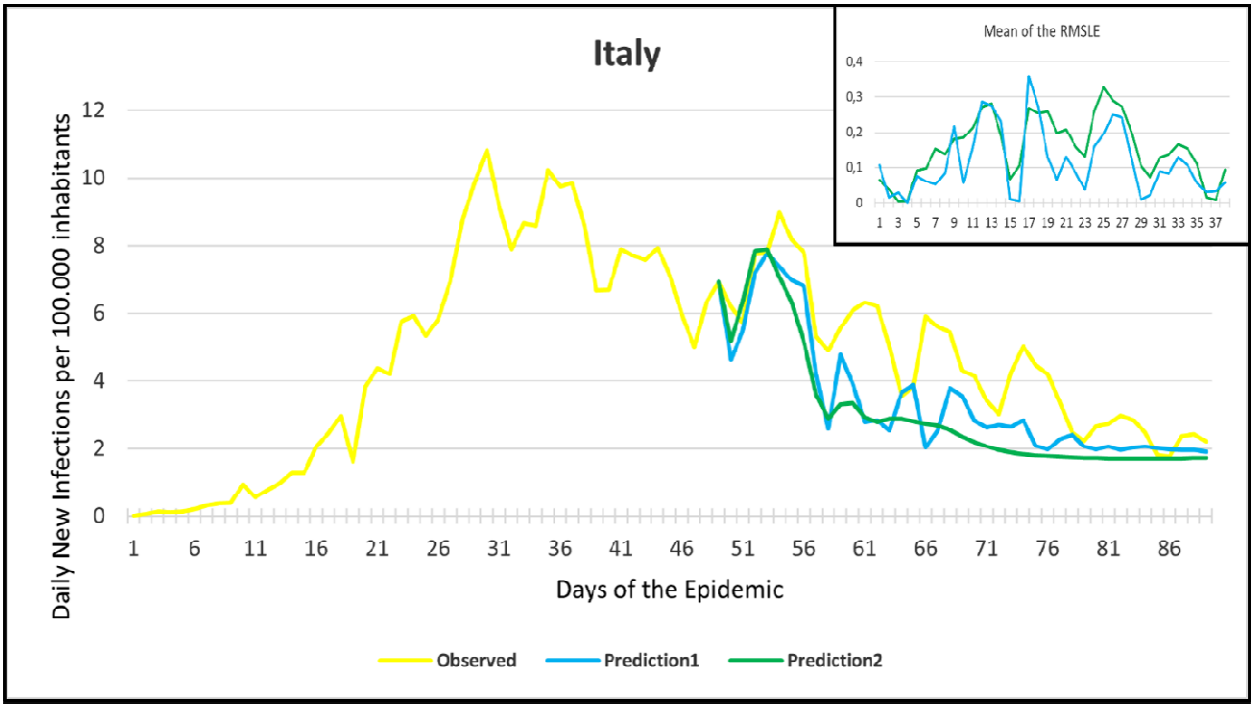

Figure 7. Observation and predictions for Italy. The small graph in the upper right corner shows the daily error values calculated for the predictions.

medRxiv preprint doi: https://doi.org/10.1101/2020.04.17.20069666; this version posted January 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

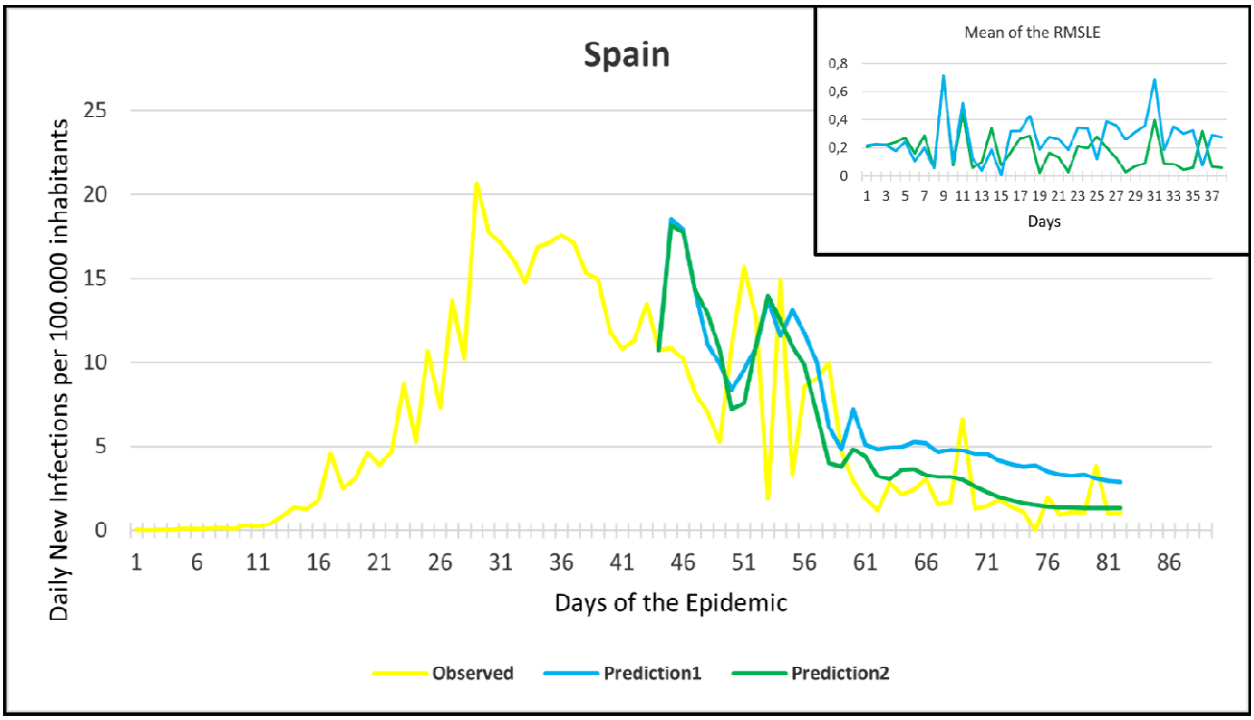

Figure 8. Observation and predictions for Spain. The small graph in the upper right corner shows the daily error values calculated for the predictions.

medRxiv preprint doi: https://doi.org/10.1101/2020.04.17.20069666; this version posted January 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

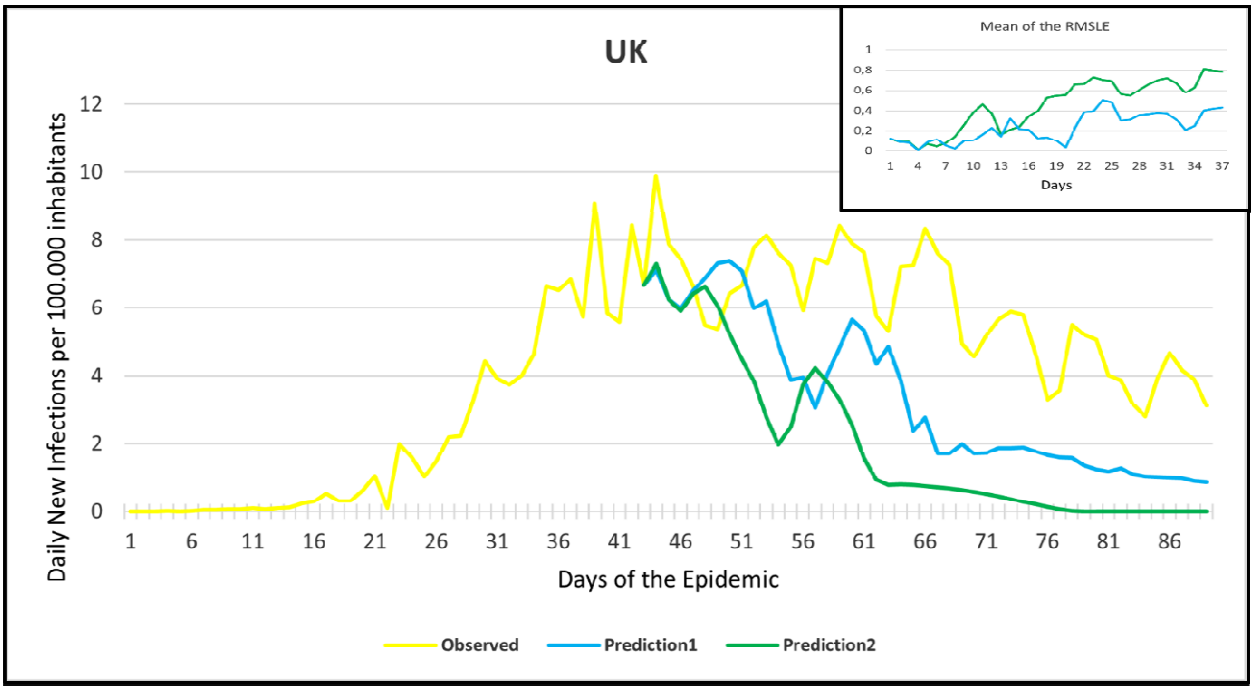

Figure 9. Observation and predictions for the United Kingdom (UK). The small graph in the upper right corner shows the daily error values calculated for the predictions.

medRxiv preprint doi: https://doi.org/10.1101/2020.04.17.20069666; this version posted January 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

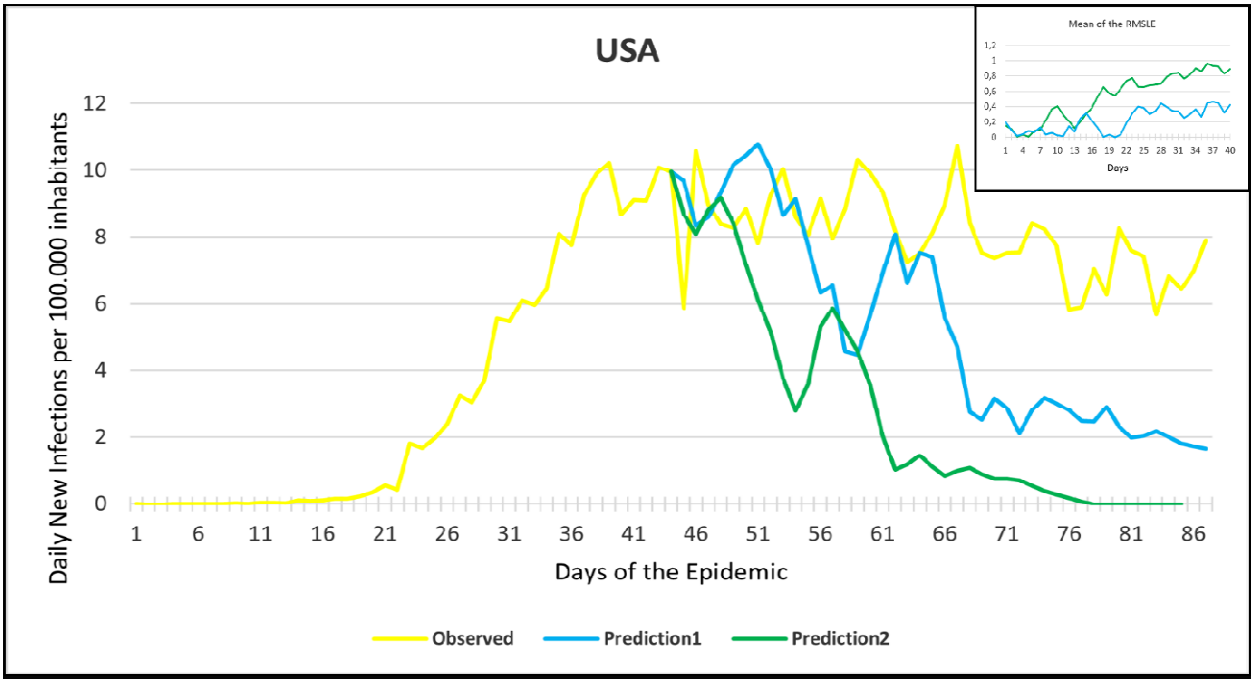

Figure 10. Observation and predictions for the United States of America (USA). The small graph in the upper right corner shows the daily error values calculated for the predictions.

medRxiv preprint doi: https://doi.org/10.1101/2020.04.17.20069666; this version posted January 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Table 1. Total RMSLE error for the entire investigated period

Country	Mean of RMSLE of Prediction 1	Mean of RMSLE of Prediction 2
Hungary	0.06	0.107
UK	0.234	0.455
Italy	0.114	0.155
Spain	0.266	0.181
Germany	0.147	0.108
France	0.513	0.307
USA	0.216	0.528

4. DISCUSSION

The current COVID-19 pandemic has opened avenues for more effective and efficient application of AI in various aspects of fighting and tackling infectious diseases. At the early stage of the COVID-19 outbreak when little data were available regarding the nature and transmissibility of SARS-CoV-2, modelling studies have attempted to predict the epidemic outcomes using Susceptible-Exposed-Infectious-Recovered (SEIR) model, based on data from Wuhan, China ; the starting point of the outbreak [29]. Besides, forecasting and predicting the COVID-19 trajectories was not the only application of AI during the current pandemic. AI has been adopted in contact-tracing , tracking public health behaviors, and currently in COVID-19 vaccination [30,31]. Since March 2020, we have had the opportunity in our present study to use the declared numbers of daily new cases of COVID-19 to predict our models and to compare the predicted trajectories with the observed data. In countries that imposed strict measures (e.g.,

medRxiv preprint doi: https://doi.org/10.1101/2020.04.17.20069666; this version posted January 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

strict lockdown), for example, France, Hungary, Italy, and Spain, the predicted models and observed data were closely similar, however, this was not the case of the UK and the USA. This could be attributed to the fact that the learning dataset was based mainly on data from Hubei province where a strict total lockdown was imposed there, unlike the UK and the USA. Besides, we can notice in the models that *Prediction 1* is more accurate for some countries (Hungary, Italy, UK, USA), while *Prediction 2* is more accurate for others (France, Germany, Spain).

The findings of our study underscores that the COVID-19 pandemic is a propagated source outbreak, therefore repeated peaks on the epidemic curve (rise of the daily number of newly diagnosed infections) are to be anticipated. Predictions that were made using AI-based RNNs further implicate that albeit the majority of investigated countries are near or over the peak of the curve, they should prepare for a series of successively high peaks in the near future, until all susceptible people will be infected by the SARS-CoV-2, or effective preventive (e.g., vaccination) or treatment options will become available and utilized effectively. These scenarios are similar to other known propagated source epidemics, e.g., SARS and measles [32].

Albeit suppression and mitigation measures can reduce the incidence of infection, COVID-19 disease, given its relatively high transmissibility reflected by average R0 values of 3.28, will continue to spread, most likely [16]. Accordingly, public health measures must be implemented as the incubation period of the virus may be long (1-14 days, but there are some opinions, that this can be 21 days), during which time asymptomatic or pre-symptomatic spreading may ensue. Moreover, currently it is uncertain, whether those, who were diagnosed with COVID-19 infection, will acquire sufficient immunity or not [13]. Finally, data from countries with warm climates suggest that summer is unlikely to stop the pandemic, as the virus already spreading in Australia and South Africa as well [15,19]. This is why the recurrence of another peak is very likely, and the end of the pandemic cannot be accurately predicted at this time.

Nevertheless, recent publications showed that the earlier the mitigation attempts are in place (e.g., border closure, closing schools, the lockdown of the country, curfew), the more effective is the reduction of the spread of the epidemic [19]. In fact, analyzing the effects of a suppression strategy concerning the COVID-19, it was shown that early implementation of suppression at 0.2 deaths per 100 000 population per week could save 30.7 million lives

medRxiv preprint doi: https://doi.org/10.1101/2020.04.17.20069666; this version posted January 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

compared to late implementation of these measures at 1.6 deaths per 100 000 population per week [33]. This seems to be the case in the countries, which had prior knowledge regarding coronavirus infections (e.g., China, Singapore, Hong Kong), as they were more prepared to implement public health measures, had more equipment and health care personnel in place to mitigate the spread of the infections. Those countries, that failed to implement efficient and strict mitigation policies in a timely manner, were facing difficulty in controlling the spread of the disease, as is the case of Italy, the UK and the USA [16].

To the best of our knowledge, this is the first study to model the predicted evolution of the newly diagnosed infections using data from official databases with the help of the AI-based RNNs trained on the currently available data, which were validated by RMSLE calculation. Most studies to date expect a single peak of the epidemic curve, but some fear the emergence of future peaks when mitigation-suppression measures will be discontinued. According to our model, this can even happen, if strict measures are sustained.

Nevertheless, the are some limitations in our study. As the nature of SARS-COV-2 is relatively unknown or dynamic, and it is prone to mutations, the prediction of the spread of the pandemic is not an easy mission. Factors that influenced the reported new cases per day, for example, the efficiency of reporting, the different quality and timing of public health measures, country-specific age-pyramid, and chronic disease burden of the population were not included in our training data set due to lack of reliable data. We did not investigate the number of deaths and recoveries, as we found no reliable data at that time. Similarly, the data regarding diagnostic tests performed per country, or death rates were omitted, given they are highly influenced by the countries' economic wellbeing, health care systems, facilities and capacities, and other factors [34,35]. There are lots of unforeseen uncertainties and coincidences which could not be implemented in our model, for example, there were days when a large number of people have been diagnosed with COVID-19 on one day (for example in care homes in France or Hungary) that caused a large increase in the number of the daily new cases [16].

To summarize, the COVID-19 disease is a global health challenge, which forced the WHO to declare it a "public health emergency of international concern on 30/01/2020" and later as a global pandemic [18]. The influence of this global epidemic has dug deep into the day-to-day conduct of everyone, with unforeseen challenges still pending for governments and

medRxiv preprint doi: https://doi.org/10.1101/2020.04.17.20069666; this version posted January 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

policymakers. Starting from this, everyone, especially decision-makers must be aware, that the current situation might be just the beginning, and even if strict public health measures are executed and sustained, future peaks of infections are possible.

5. CONCLUSION

The findings of our study underscore that the COVID-19 pandemic is a propagated source epidemic, therefore repeated peaks of the rise of the daily number of newly diagnosed infections are to be anticipated. In countries where strict control measures were imposed, the predicted models were closely similar to the observed data. Most studies to date expect a single peak on the epidemic curve, but some fear the emergence of future peaks when mitigation-suppression measures will be discontinued. According to our models, this can even happen, if strict measures are sustained. The AI-based predictions might be useful tools and can be recalculated according to the newly observed data to get a more precise forecast of the pandemic. Finally, AI-based predictions are expected to provide public health practitioners and decision makers with sufficient data that would be useful in improving countries' preparedness to the next stage of a pandemic.

medRxiv preprint doi: https://doi.org/10.1101/2020.04.17.20069666; this version posted January 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Author Contributions: Conceptualization, LRK, TB, JZ, AH; methodology, the methodology was discussed by all co-authors; Data curation: AH, ABA, TB, AT, IV led the data curation and formal analysis; TB, ABA, and GJS made the data extraction and updating. The AI-based analysis and validation was made by AT, TB, IV, AH. Visualizations were made by AT, TB, IV, GJS, AH, ABA, and LRK. Writing—original draft preparation, all authors; Writing—review and editing, all authors; funding acquisition, JZ, and AH. The literature search was done by LRK, ABA, AH, SH, GJS, SG, JZ, RG. The interpretation and review of the literature, methods, results were made in close collaboration by all co-authors. All authors have read and agreed to the submitted version of the manuscript.

Funding: This study was supported by the European Union, co-financed by the European Social Fund and European Regional Development Fund [Grant No. EFOP-3.6.1-16-2016-00022 Debrecen Venture Catapult Program (providing support for SH, LRK), by the European Union [EFOP-3.6.2-16-2017-00009 Establishing Thematic Scientific and Cooperation Network for Clinical Research (providing support for RG), by the Janos Bolyai Research Fellowship of the Hungarian Academy of Sciences (providing support for JZ), by the Hungarian Brain Research Program 2.0 under grant number 2017-1.2.1-NKP-2017-00002 and the ED_18-1-2019-0028 providing support for (LRK, TB, AT). The research was also supported by the UNKP-19-3 – I. New National Excellence Program of the Ministry for Innovation and Technology (AT). The research was supported in part by the project EFOP-3.6.2-16-2017-00015 supported by the European Union, co-financed by the European Social Fund. This work was also supported in part by the project EFOP-3.6.3-VEKOP-16-2017-00002, sup-ported by the European Union, co-financed by the European Social Fund (IV). The funders had no role in the writing of the manuscript or the decision to submit it for publication, no involvement in data collection, analysis, or interpretation; trial design; patient recruitment; or any aspect pertinent to the study.

Institutional Review Board Statement: Not Applicable.

Informed Consent Statement: Not Applicable

Data Availability Statement: All data sources are publicly available and were described and cited in the methods section.

medRxiv preprint doi: https://doi.org/10.1101/2020.04.17.20069666; this version posted January 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Acknowledgments: ABA is supported by the Tempus Public Foundation's Scholarship at the University of Debrecen. Also, the authors would like to provide their thanks and appreciation to the World Health Organization and Johns Hopkins University for the publicly available data.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

REFERENCES

1. Han, Q.; Lin, Q.; Jin, S.; You, L. Coronavirus 2019-nCoV: A brief perspective from the front line. *J. Infect.* **2020**, *80*, 373–377, doi:10.1016/j.jinf.2020.02.010.

2. Ksiazek, T.G.; Erdman, D.; Goldsmith, C.S.; Zaki, S.R.; Peret, T.; Emery, S.; Tong, S.; Urbani, C.; Comer, J.A.; Lim, W.; et al. A Novel Coronavirus Associated with Severe Acute Respiratory Syndrome. *N. Engl. J. Med.* **2003**, *348*, 1953–1966, doi:10.1056/NEJMoa030781.

3. CHAN-YEUNG, M.; XU, R.-H. SARS: epidemiology. *Respirology* **2003**, *8*, S9–S14, doi:https://doi.org/10.1046/j.1440-1843.2003.00518.x.

4. Cheng, V.C.C.; Lau, S.K.P.; Woo, P.C.Y.; Yuen, K.Y. Severe Acute Respiratory Syndrome Coronavirus as an Agent of Emerging and Reemerging Infection. *Clin. Microbiol. Rev.* **2007**, *20*, 660–694, doi:10.1128/CMR.00023-07.

5. Peiris, J.S.M.; Lai, S.T.; Poon, L.L.M.; Guan, Y.; Yam, L.Y.C.; Lim, W.; Nicholls, J.; Yee, W.K.S.; Yan, W.W.; Cheung, M.T.; et al. Coronavirus as a possible cause of severe acute respiratory syndrome. *Lancet* **2003**, *361*, 1319–1325, doi:https://doi.org/10.1016/S0140-6736(03)13077-2.

6. de Groot, R.J.; Baker, S.C.; Baric, R.S.; Brown, C.S.; Drosten, C.; Enjuanes, L.; Fouchier, R.A.M.; Galiano, M.; Gorbalenya, A.E.; Memish, Z.A.; et al. Commentary: Middle East Respiratory Syndrome Coronavirus (MERS-CoV): Announcement of the Coronavirus Study Group. *J. Virol.* **2013**, *87*, 7790–7792, doi:10.1128/JVI.01244-13.

medRxiv preprint doi: https://doi.org/10.1101/2020.04.17.20069666; this version posted January 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

7. World Health Organization MERS Situation Update Available online: http://www.emro.who.int/pandemic-epidemic-diseases/mers-cov/mers-situation-update-january-2020.html (accessed on Mar 29, 2020).

8. Rabi, F.A.; Al Zoubi, M.S.; Al-Nasser, A.D.; Kasasbeh, G.A.; Salameh, D.M. Sars-cov-2 and coronavirus disease 2019: What we know so far. *Pathogens* **2020**, *9*, 231, doi:10.3390/pathogens9030231.

9. Al-Tammemi, A.B. The Battle Against COVID-19 in Jordan□: An Early Overview of the Jordanian Experience. *Front. Public Heal.* **2020**, *8*, 188, doi:10.3389/fpubh.2020.00188.

10. Zhou, P.; Yang, X.-L.; Wang, X.-G.; Hu, B.; Zhang, L.; Zhang, W.; Si, H.-R.; Zhu, Y.; Li, B.; Huang, C.-L.; et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **2020**, *579*, 270–273, doi:10.1038/s41586-020-2012-7.

11. Chen, N.; Zhou, M.; Dong, X.; Qu, J.; Gong, F.; Han, Y.; Qiu, Y.; Wang, J.; Liu, Y.; Wei, Y.; et al. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. *Lancet* **2020**, *395*, 507–513, doi:https://doi.org/10.1016/S0140-6736(20)30211-7.

12. Sun, P.; Ren, J.; Li, K.; Qie, S.; Liu, Z.; Xi, J. Response to: Sore throat in COVID-19: comment on "Clinical characteristics of hospitalized patients with SARS-CoV-2 infection: A single arm meta-analysis." *J. Med. Virol.* **2020**, *92*, 716–718, doi:https://doi.org/10.1002/jmv.25818.

13. World Health Organization Coronavirus disease (COVID-19) - Q&A Available online: https://www.who.int/news-room/q-a-detail/q-a-coronaviruses (accessed on Apr 10, 2020).

14. Wu, J.; Zha, P. Association of COVID-19 Disease Severity with Transmission Routes and Suggested Changes to Community Guidelines. *SSRN Prepr.* **2020**, doi:10.20944/preprints202003.0246.v1.

15. Yuen, K.-S.; Ye, Z.-W.; Fung, S.-Y.; Chan, C.-P.; Jin, D.-Y. SARS-CoV-2 and COVID-19: The most important research questions. *Cell Biosci.* **2020**, *10*, 40, doi:10.1186/s13578-020-00404-4.

medRxiv preprint doi: https://doi.org/10.1101/2020.04.17.20069666; this version posted January 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

16. World Health Organization Coronavirus disease (COVID-19) Weekly Epidemiological Update and Weekly Operational Update Available online: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports (accessed on Apr 15, 2020).

17. Liu, Y.; Gayle, A.A.; Wilder-Smith, A.; Rocklöv, J. The reproductive number of COVID-19 is higher compared to SARS coronavirus. *J. Travel Med.* **2020**, *27*, doi:10.1093/jtm/taaa021.

18. Wu, Z.; McGoogan, J.M. Characteristics of and Important Lessons From the Coronavirus Disease 2019 (COVID-19) Outbreak in China: Summary of a Report of 72□314 Cases From the Chinese Center for Disease Control and Prevention. *JAMA* **2020**, *323*, 1239–1242, doi:10.1001/jama.2020.2648.

19. Ferguson, N.; Laydon, D.; Nedjati Gilani, G.; Imai, N.; Ainslie, K.; Baguelin, M.; Bhatia, S.; Boonyasiri, A.; Cucunuba Perez, Z.; Cuomo-Dannenburg, G. Report 9: Impact of non-pharmaceutical interventions (NPIs) to reduce COVID19 mortality and healthcare demand Available online: https://www.imperial.ac.uk/media/imperial-college/medicine/sph/ide/gida-fellowships/Imperial-College-COVID19-NPI-modelling-16-03-2020.pdf (accessed on Mar 16, 2020).

20. Hethcote, H.W. *Modeling heterogeneous mixing in infectious disease dynamics*; Cambridge University Press Cambridge, 1996; Vol. 6;.

21. Siettos, C.I.; Russo, L. Mathematical modeling of infectious disease dynamics. *Virulence* **2013**, *4*, 295–306, doi:10.4161/viru.24041.

22. Szolovits, P. *Artificial intelligence in medicine*; Routledge, 2019; ISBN 0429728468.

23. Zhu, X.; Fu, B.; Yang, Y.; Ma, Y.; Hao, J.; Chen, S.; Liu, S.; Li, T.; Liu, S.; Guo, W.; et al. Attention-based recurrent neural network for influenza epidemic prediction. *BMC Bioinformatics* **2019**, *20*, 575, doi:10.1186/s12859-019-3131-8.

24. Dupond, S. A thorough review on the current advance of neural network structures. *Annu. Rev. Control* **2019**, *14*, 200–230.

medRxiv preprint doi: https://doi.org/10.1101/2020.04.17.20069666; this version posted January 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

25. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780, doi:10.1162/neco.1997.9.8.1735.

26. Johns Hopkins University NEW CASES OF COVID-19 IN WORLD COUNTRIES Available online: https://coronavirus.jhu.edu/data/new-cases (accessed on Apr 12, 2020).

27. Yoon, S.; Yun, H.; Kim, Y.; Park, G.; Jung, K. Efficient transfer learning schemes for personalized language modeling using recurrent neural network. *arXiv Prepr.* **2017**, doi:https://arxiv.org/abs/1701.03578.

28. Sutskever, I.; Vinyals, O.; Le, Q. V Sequence to Sequence Learning with Neural Networks Available online: https://arxiv.org/abs/1409.3215v3 (accessed on Apr 12, 2020).

29. Ogden, N.H.; Fazil, A.; Arino, J.; Berthiaume, P.; Fisman, D.N.; Greer, A.L.; Ludwig, A.; Ng, V.; Tuite, A.R.; Turgeon, P.; et al. Modelling scenarios of the epidemic of COVID-19 in Canada. *Can. Commun. Dis. Rep.* **2020**, *46*, 198–204, doi:10.14745/ccdr.v46i06a08.

30. Ganasegeran, K.; Abdulrahman, S.A. Artificial Intelligence Applications in Tracking Health Behaviors During Disease Epidemics. In *Human Behaviour Analysis Using Intelligent Systems*; Hemanth, D.J., Ed.; Springer International Publishing: Cham, 2020; pp. 141–155 ISBN 978-3-030-35139-7.

31. Lalmuanawma, S.; Hussain, J.; Chhakchhuak, L. Applications of machine learning and artificial intelligence for Covid-19 (SARS-CoV-2) pandemic: A review. *Chaos, Solitons & Fractals* **2020**, *139*, 110059, doi:https://doi.org/10.1016/j.chaos.2020.110059.

32. Centers for Disease Control and Prevention Using an Epi Curve to Determine Mode of Spread. Available online: https://www.cdc.gov/training/QuickLearns/epimode/6.html (accessed on Apr 12, 2020).

33. Walker, P.; Whittaker, C.; Watson, O.; Baguelin, M.; Ainslie, K.; Bhatia, S.; Bhatt, S.; Boonyasiri, A.; Boyd, O.; Cattarino, L. Report 12: The global impact of COVID-19 and strategies for mitigation and suppression Available online: https://www.imperial.ac.uk/mrc-global-infectious-disease-analysis/covid-19/report-12-global-impact-covid-19/ (accessed on Apr 12, 2020).

medRxiv preprint doi: https://doi.org/10.1101/2020.04.17.20069666; this version posted January 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

34. Petropoulos, F.; Makridakis, S. Forecasting the novel coronavirus COVID-19. *PLoS One* **2020**, *15*, e0231236.

35. Anastassopoulou, C.; Russo, L.; Tsakris, A.; Siettos, C. Data-based analysis, modelling and forecasting of the COVID-19 outbreak. *PLoS One* **2020**, *15*, e0230405.