# The Epidemic Volatility Index: an early warning tool for epidemics

Polychronis Kostoulas[1]

[1]Faculty of Public Health, University of Thessaly, Greece. Email: pkost@uth.gr

April 23, 2021

**Abstract**

The objective of this work is to present the Epidemic Volatility Index (EVI), an early warning tool for upcoming epidemic waves. EVI is based on the volatility of the newly reported cases per unit of time, ideally per day, and issues an early warning if the rate of the volatility change exceeds a threshold. EVI is simple and its application on data from the current COVID-19 pandemic, revealed a consistent and stable performance in predicting the waves of the COVID-19 epidemic for each one of the world countries. The application of EVI in the case of other epidemics and syndromic surveillance is straightforward and its combination with existing alarm systems will promote the early implementation of appropriate interventions and the successful containment of outbreaks.

## Introduction

Early warning tools are crucial for the timely application of intervention strategies that will effectively control the spread of the etiologic agent and mitigate the adverse health, social and economic effects of an epidemic. Early warning systems can be based on the seasonality of epidemics and the link between pathogens and meteorological parameters [1] and/or the measurement of vector indices for vector-borne pathogens[2]. Further, sentinel networks in combination with information technology infrastructures in public health[3] provide data for the detection of spatial and temporal aberrations in the expected number of cases for groups of signs and symptoms. Several modelling frameworks exist for the analysis of such data, as, for example, the moving epidemic method, an approach used to monitor, among others, the start of the flu epidemic[4].

Once an epidemic erupts, growth models - mainly based on $R_0$ - can be used to predict the course of the outbreak and quantify its consequences. The advantages and limitations of these methods have been extensively discussed [5]. Machine learning algorithms have also been utilized with the most recent application being in the current COVID-19 pandemic[6]. Correlation of parameters, from big data sources, with the number of COVID-19 cases has been used to predict a future rise in cases. For example, monitoring of digital data streams can be an early sign for a rise in the COVID-19 cases and deaths in the next 2-3 weeks[7]. All models have limitations arising from the imperfect nature of the data. The need for open, better, detailed data is imperative for the deployment of models with improved accuracy, models that will have better predictive ability and will be more useful for the timely application of appropriate control measures for the COVID-19 pandemic[8].

In this work, the Epidemic Volatility Index ($EVI$) is presented for the first time. $EVI$ has been inspired by the use of volatility indices in the stock market. Volatility has mainly a negative association with stocks or stocks' future prices [9,10]. $EVI$ is based on the moving standard deviation of the newly reported cases during an epidemic. Initially, $EVI$ is presented and then an example application is given with COVID-19 data. Results revealed a firm and consistent ability of $EVI$ to predict the waves of the COVID-19 epidemic in each one of the world countries.

# Methods

## The Epidemiologic Volatility Index

The epidemiologic volatility index is calculated for a rolling window of time series epidemic data (i.e. the number of new cases per day). At each step, the observations within the window that are analyzed are obtained by shifting the window forward over the time series data one observation at a time. Let $y_i = \{y_1,\ y_2, ..., y_n\}$ be a time series of length $N$. The rolling window size - that is the number of consecutive observations per rolling window - is $m$. With $0 < m \leq m_{\max}$ and $0 \leq m_{\max} \leq N$, there are $t = N - m + 1$ consecutive rolling windows.

At each of the $t$ steps, $EVI$ uses the standard deviation $(\sigma_t)$ of the newly reported cases $(y_{j_t} = \{y_{1_t},\ y_{2_t}, \dots, y_{m_t}\})$ within the specified $m$:

$$\sigma_t = \sqrt{\frac{1}{m}\sum_{i_t=1}^{m}(x_{i_t} - \overline{\mu}_t)}$$

with $\overline{\mu_t}$ the mean of the $t^{th}$ window. Subsequently, $EVI$ is calculated as the relative change of $(\sigma_t)$ between two consecutive rolling windows:

$$EVI_{t-1,t} = \frac{\sigma_t - \sigma_{t-1}}{\sigma_t}$$

We expect an increase in the future number of cases, if $EVI_{t-1,t}$ exceeds a threshold $c$ $(c \in [0,1])$ and the observed cases at time point $t, (y_t)$ are higher than the average of the reported cases in the previous week:

$$Ind_{EVI_{t-1,t}} = \begin{cases} 1 & if \quad EVI_{t-1,t} >= c \ \ \wedge \ \ y_t >= \overline{\mu}_{t:t-7} \\ 0 & \text{otherwise} \end{cases}$$

## Case definition and desired accuracy

The user should provide a case definition of what constitutes the minimum expected rise in cases that, if present, should be detected. A case definition can be a rise in cases between two consecutive weeks that exceeds a threshold $r$:

$$\frac{\overline{\mu}_{t:t-7}}{\overline{\mu}_{t:t+7}} \geq r$$

with $0 \leq r \leq 1$.

The accuracy of $EVI$, given the specified case definition, depends on $m$ and $c$, which should be selected in a way to achieve a desired accuracy target. Several strategies are available. Selection of $m$ and $c$ values that lead to the simultaneous maximization of the sensitivity $(Se)$ and the specificity $(Sp)$ for $EVI$ and the Youden index $(J = Se + Sp - 1)$[11] and thus to an overall minimization of the false results (i.e. both false positive and false negative early warnings). Another approach could be to select $m$ and $c$ such that the highest $Se\,(or\ Sp)$ is achieved with $Sp\,(or\ Se) = 1$ not dropping below a critical value (e.g. 95%). Advanced Receiver Operating Characteristic curve analysis can also be performed [12] and selection of critical values can be based on indices that quantify the relative cost of false positive (i.e., falsely predicting an upcoming

epidemic wave) to false negative (i.e., failing to predict an upcoming epidemic wave) warnings, like the misclassification cost term ($MCT$).

## Issuance of an early warning

Every time a new time point $t$ is observed, the model uses all the observed cases up to $t$ to decide whether it should issue an early warning, at time point $t$. The steps are:

1. Observed cases up to $t$ are analyzed for all possible values for the window size ($m \in [1, m_{\max}]$) and threshold ($c \in [0, 1]$).
2. For each of the $m, c$ combinations the $Se_{t_{m,c}}$and $Sp_{t_{m,c}}$is estimated for the predefined case definition (Eq. 4).
3. The $m'$ and $c'$ that give the best $Se_{t_{m',c'}}$ and $Sp_{t_{m',c'}}$ combination are selected.
4. For $m'$ and $c'$, the value of $Ind_{EVI_{t,t-1}}$ is determined at the most recent time point $t$ and a decision is made on whether a warning signal is issued.

## Accuracy, Positive and Negative Predictive Values

Further, at each time point $t$, the probability of observing a rise or drop in the future cases given that an early warning was issued or not can be calculated as the positive ($PV_t+$) and negative ($PV_t-$) predictive value, respectively:

$$PV_t+ = P(D+ \mid T+) = \frac{p_{1:t} Se_{t_{m',c'}}}{p_{1:t} Se_{t_{m',c'}} + (1 - p_{1:t}) \left(1 - Sp_{t_{m',c'}}\right)}$$

$$PV_t- = P(D- \mid T-) = \frac{(1 - p_{1:t})\, Sp_{t_{m',c'}}}{(1 - p_{1:t})\, Sp_{t_{m',c'}} + p_{1:t} \left(1 - Se_{t_{m',c'}}\right)}$$

where $p_{1:t}$ is the proportion of events satisfying the condition of Eq. 4 up to time point $t$.

Once the entire time series data has been observed, the overall $Se$ of $EVI$ can be estimated as the fraction of the total number of occurrences for which an early warning was issued, given that the case definition (Eq. 4 ) holds ($P(T+ \mid D+)$), divided by the total number of occurrences that the case definition holds ($P(D+)$). Similarly, the overall $Sp$ of $EVI$ is calculated as the fraction of the total number of occurrences for which an early warning was not issued given that the expected rise of cases was not observed, that is, the case definition is not true, ($P(T- \mid D-)$) divided by the total number of occurrences that the case definition is not true ($P\,(D-)$):

$$Se = \frac{P\,(T+ \mid D+)}{P\,(D+)},\ Sp = \frac{P\,(T- \mid D-)}{P\,(D-)}$$

## Sensitivity analysis

The performance of EVI depends on the specified case definition (i.e., $r$) and the desired accuracy. Ideally, in the presence of historical data, various case definitions and $r$ values should be explored to identify combinations that provide the optimal monitoring of an epidemic.

### Example application

The current most serious threat to global health and economy[13] is the COVID-19 pandemic that begun in China and was first reported to the WHO China Country Office on December 31, 2019[14]. Data on the confirmed cases of COVID-19 were retrieved by the COVID-19 Data Repository, which is maintained by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University[15]. The number of daily confirmed new cases of COVID-19, for each country, from January 22, 2020 until April 13, 2021 were analyzed. Due to unnatural variability in the reporting cases between working days and weekends, the 7-day moving average rather than the actually observed cases were analyzed. For the analysis $m_{max}$ was restricted to 30 days in order to avoid the effect of potentially higher volatility from previous epidemic waves on the volatility estimates of the most recent data and the predicitve ability of $EVI$ for upcoming and perhaps milder epidemic waves.

The case definition was an increase in the mean of expected cases, between two consecutive weeks, equal or higher than twenty percent, $r \geq \frac{1}{1.2}$. For sensitivity analysis, the detection of an increase in the mean of expected cases equal or higher than 50 percent ($r \geq \frac{1}{1.5}$) was considered. Data were analyzed separately for each country and for each of the United States of America that had, until April 13, 2021, experienced a total number of cases higher than 20,000.

### Statistical software

All models were run in R[16] and the packages readxl[17], ggplot2[18], cowplot[19] and readr[20] were used.

## Results

Results for Italy, one of the most severely affected EU countries[21], and New York, which was in the epicenter of the pandemic in the U.S.[22], are presented in the main manuscript. Results from a list of selected countries, namely, Argentina, Australia, Belgium, Brazil, California, Canada, Czechia, Florida, France, Germany, Greece, India, Italy, Netherlands, New York, Poland, Portugal, Romania, Russia, Saudi Arabia, South Africa, Spain, Sweden, Texas Ukraine and the United Kingdom, are provided as a supplementary material along with results from all countries in the world and the Unites States (S1, S2 and S3, respectively).

Confirmed cases COVID-19 for Italy and New York, from January 22, 2020 until April 13, 2021, are in Figures 1 and 2, respectively. Red dots correspond to time points that an early warning was issued according to $Ind_{EVI_{t,t-1}}$, while grey dots to time points without an early warning indication. Further, the positive and negative predictive values at each time point are in Figures 3 and 4, respectively.

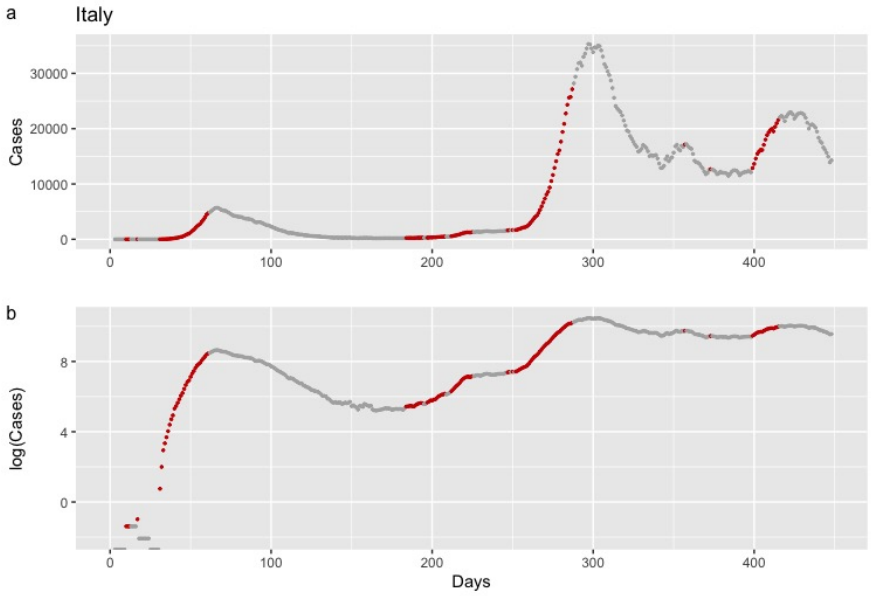


Figure 1: Daily confirmed cases of COVID-19 in Italy, January 22, 2020 until April 13, 2021. Red dots correspond to dates that, according to the Epidemic Volatility Index (EVI), an early warning was issued indicating that a rise in the COVID-19 cases is expected. Data are presented on the original scale (1a) and the logarithmic scale (1b), which facilitates the comparison of the steepness of the epidemic curve between the different waves of the pandemic.

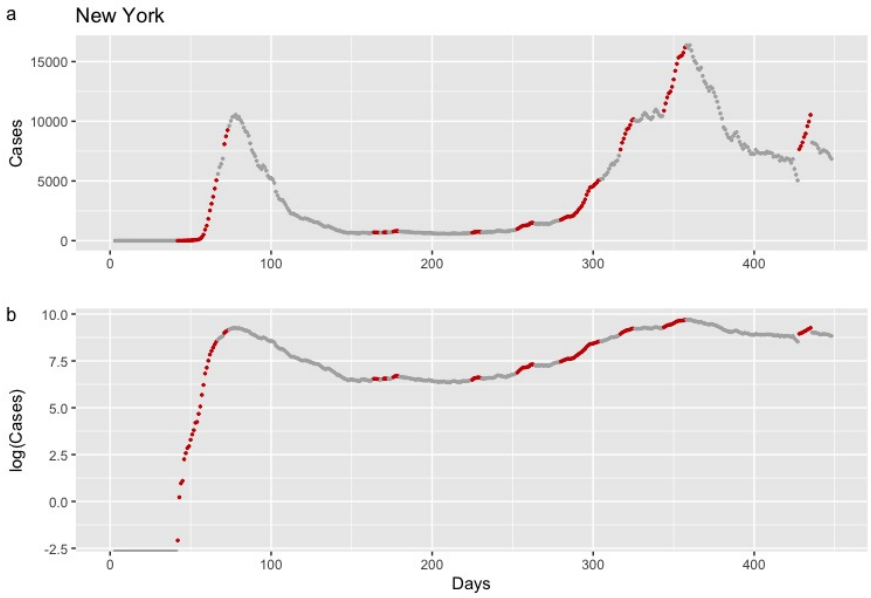


Figure 2: Daily confirmed cases of COVID-19 in Italy, January 22, 2020 until April 13, 2021. Red dots correspond to dates that, according to EVI, an early warning was issued indicating that a rise in the COVID-19 cases is expected. Data are presented on the original scale (1a) and the logarithmic scale (1b), which facilitates the comparison of the steepness of the epidemic curve between the different waves of the pandemic.

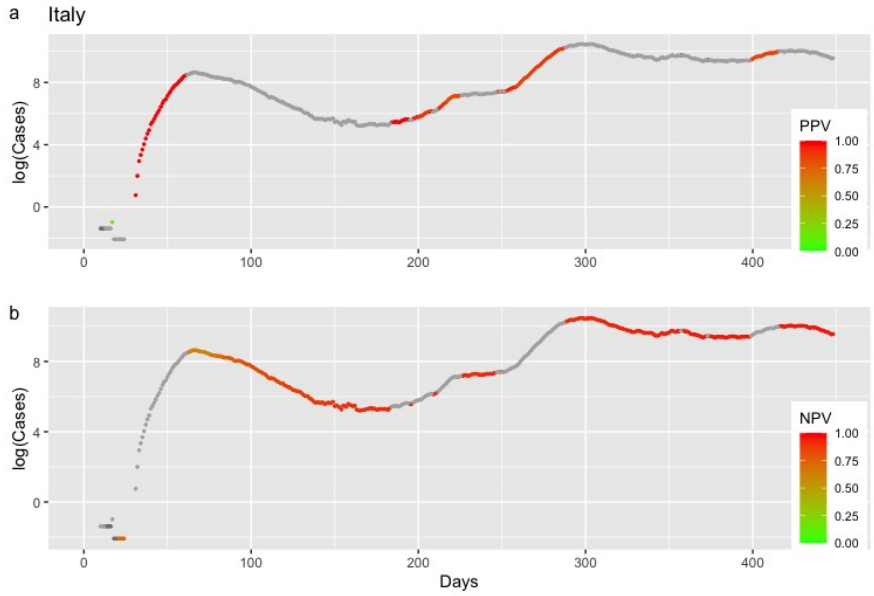


Figure 3: Positive and negative predictive values (PPV in 3a and NPV in 3b), for Italy, depending on whether an early warning was isseud or not. Higher color intensity corresponds to predicitve values closer to the value of 1.

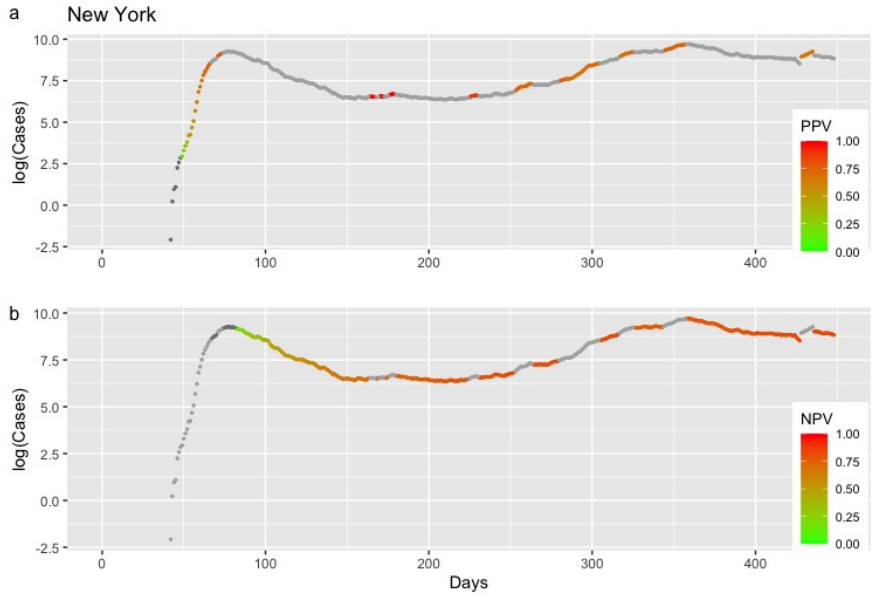


Figure 4: Positive and negative predictive values (PPV in 4a and NPV in 4b), for Italy, depending on whether an early warning was isseud or not. Higher color intensity corresponds to predicitve values closer to the value of 1.

For Italy, the overall sensitivity for EVI was 0.82 and the specificity was 0.91 (0.88; 0.94). For New York, the corresponding values were 0.55 (0.47; 0.64) and 0.88 (0.84; 0.91).

Sensitivity analysis results for Italy, under alternative $r$ specification (i.e., $r = \frac{1}{1.5}$), are in Figure 5. The overall sensitivity and specificity, for $r = \frac{1}{1.5}$, was 0.75 (0.66; 0.85) and 0.93 (0.91; 0.96).

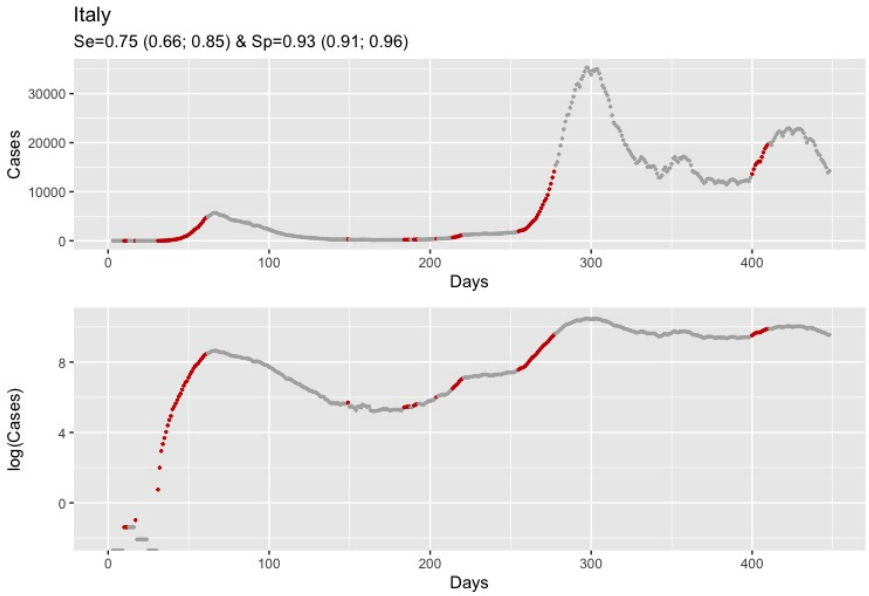


Figure 5: Daily confirmed cases of COVID-19 in Italy, January 22, 2020 until April 13, 2021. Analysis with $r = \frac{1}{1.5}$. Red dots correspond to dates that, based on EVI, an early warning was issued indicating that a rise in the COVID-19 cases is expected. Data are presented on the original scale (1a) and the logarithmic scale (1b) which facilitates the comparison of the steepness of the epidemic curve between the different waves of the pandemic.

A common and consistent finding in the results from all countries was that repetitive early warnings are linked to the beginning of a new epidemic wave, while absence of warnings indicate a stable course or a future drop in the number of COVID-19 new cases (Figure 1 and 2 and supplementary files S1, S2 and S3).

# Discussion

$EVI$ is an efficient, early-warning tool for an upcoming rise in the number of new cases. This ability of $EVI$, as expressed by its overall $Se$ and $Sp$, was, in all instances, high. A more important aspect lies in the fact that repetitive issuance of early warnings indicates the beginning of future epidemic waves. This was a consistent and remarkably stable finding across all countries (Fig. 1 and 2, supplementary files S1, S2 and S3). In a similar manner, absence of a series of early warnings imply that the number of cases will remain the same or drop. The latter was also a consistent finding across all countries. Additionally, false early warnings (i.e. false positives) were isolated instances and did not occur in a consecutive series. There were few instances with a consecutive absence of early warnings despite a continuing rise in the number of cases (i.e. false negatives). Nevertheless, such series of false negatives never occurred at the beginning of an epidemic wave but was always close to the peak of the wave. Since $EVI$ depends on volatility and the rate of increase in the number of cases drops when approaching the peak of an epidemic wave, this finding is reasonable and could be interpreted as an early sign of reaching the peak. Positive and negative predictive values that are calculated at each time point can also be used to assess the probability that an early warning or its absence is true. In all instances, predictive values were high with the exception of few instances at the

beginning of the time series due to the absence of enough data.

The ability of $EVI$ to provide valid predictions does not seem to be affected by the fact that sampling and testing schemes for COVID-19 are mainly based on passive surveillance systems. $EVI$ performed equally well among different countries with different control strategies, testing intensity and reporting accuracy and despite the fact that within countries sampling and testing has changed over time and/or differs between regions [23,24].

The performance of $EVI$ depends on the specified case definition and $r$, parameters which are epidemic-specific and country-specific. Modifications to allow for different case definition and $r$ for the different periods of an epidemic are rather straightforward. Parameters $c$ and $m$ are allowed to vary and take values that would satisfy the conditions set by the defined case and the desired accuracy. A point of concern is the selection of $m_{\max}$. For an ongoing epidemic with multiple waves, as is the case with COVID-19, $m_{\max}$ should be limited to a period shorter than the whole observation period. This prevents excess volatility of past epidemic waves from affecting the most recent volatility estimates and the ability of $EVI$ to warn for upcoming waves that may be smaller and of lower volatility than previous ones. In our example, we limited $m_{\max}$ to one month. $EVI$ also depends on data quality. Detailed data at the lowest time unit (i.e., days rather than weeks) is preferable. In the COVID-19 example the 7-day moving average was analyzed instead of the daily reported cases because daily data had unnatural variability due to reporting variations between working days and weekends. Nevertheless, analysis based on the daily reported cases provided similar results (data not shown here).

Beyond the case of epidemics or rare events, like the COVID-19 pandemic, an important application of EVI can be in the context of syndromic surveillance[25], not limited to outbreaks from terrorist attacks, but in its broader sense: the detection of temporal and spatial aberrations in the expected number of cases for sign and symptom categories. Such systems already exist and utilize state-of-the-art information technologies in public health[3]. $EVI$ can provide an additional early warning tool for these systems. The presence a similar mechanism [26] would have warned Chinese authorities - or the authorities of any country - for an upcoming unusual surge in the number of cases and could have led to early control measures for COVID-19.

# Supplementary Material

In all supplementary files one country is presented per page. Each page is organized as follows:

Title: Country

Subtitle: Overall $Se$ and $Sp$

Figure a: Daily confirmed cases of COVID-19, presented on the original scale, with red dots corresponding to dates that, according to EVI, an early warning was issued.

Figure b: Daily confirmed cases of COVID-19, presented on the logarithmic scale, which facilitates the comparison of the steepness of the epidemic curve between the different waves of the pandemic. Red dots correspond to dates that, according to EVI, an early warning was issued.

Figure c: Positive predictive values (PPV ) for the days that an early warning was issued. Higher color intensity corresponds to PPV closer to the value of 1.

Figure d: Negative predictive values (NPV) for the days that an early warning was not issued. Higher color intensity corresponds to NPV closer to the value of 1.

## Supplementary file 1: Results from a list of countries

**Hosted file**

All.pdf available at https://authorea.com/users/155758/articles/516426-the-epidemic-volatility-index-an-early-warning-tool-for-epidemics

### Supplementary file 2: Results from all world countries

#### Hosted file

1_All.pdf available at https://authorea.com/users/155758/articles/516426-the-epidemic-volatility-index-an-early-warning-tool-for-epidemics

### Supplementary file 3: Results from each of the United States

#### Hosted file

1_All_US.pdf available at https://authorea.com/users/155758/articles/516426-the-epidemic-volatility-index-an-early-warning-tool-for-epidemics

## References


1.Abeku, T. A. *et al.*. Malaria epidemic early warning and detection in African highlands. *Trends in parasitology* **20**, 400–405 (2004).

2.Chang, F.-S. *et al.*. Re-assess vector indices threshold as an early warning tool for predicting dengue epidemic in a dengue non-endemic country. *PLoS Negl Trop Dis* **9**, e0004043 (2015).

3.Heffernan, R. *et al.*. Syndromic surveillance in public health practice, New York City. (2004).

4.Vega, T. *et al.*. Influenza surveillance in Europe: establishing epidemic thresholds by the moving epidemic method. *Influenza and other respiratory viruses* **7**, 546–558 (2013).

5.Chowell, G., Sattenspiel, L., Bansal, S. & Viboud, C. Mathematical models to characterize early epidemic growth: A review. *Physics of life reviews* **18**, 66–97 (2016).

6.Wang, P., Zheng, X., Li, J. & Zhu, B. Prediction of epidemic trends in COVID-19 with logistic model and machine learning technics. *Chaos, Solitons & Fractals* **139**, 110058 (2020).

7.Kogan, N. E. *et al.*. An early warning approach to monitor COVID-19 activity with multiple digital traces in near real time.. *Sci Adv* **7**, (2021).

8.Vespignani, A. *et al.*. Modelling COVID-19. *Nature Reviews Physics* **2**, 279–281 (2020).

9.Fernandes, M., Medeiros, M. C. & Scharth, M. Modeling and predicting the CBOE market volatility index. *Journal of Banking & Finance* **40**, 1–10 (2014).

10.Brenner, M. & Galai, D. New Financial Instruments for Hedge Changes in Volatility. *Financial Analysts Journal* **45**, 61–65 (1989).

11.Fluss, R., Faraggi, D. & Reiser, B. Estimation of the Youden Index and its Associated Cutoff Point. *Biometrical Journal* **47**, 458–472 (2005).

12.Zweig, M. H. & Campbell, G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clinical Chemistry* **39**, 561–577 (1993).

13.Fauci, A. S., Lane, H. C. & Redfield, R. R. Covid-19 — Navigating the Uncharted. *New England Journal of Medicine* **382**, 1268–1269 (2020).

14.Organization, W. H. & others. Pneumonia of unknown cause—China: disease outbreak news. Geneva; January 5. (2020).

15.Dong, E., Du, H. & Gardner, L. An interactive web-based dashboard to track COVID-19 in real time.. *Lancet Infect Dis* **20**, 533–534 (2020).

16.Team, R. C. & others. R: A language and environment for statistical computing. (2020).

17.Wickham, H. *et al.*. Package 'readxl'. (2019).

18.Wickham, H. ggplot2. *Wiley Interdisciplinary Reviews: Computational Statistics* **3**, 180–185 (2011).

19.Wilke, C. O., Wickham, H. & Wilke, M. C. O. Package 'cowplot'. *Streamlined Plot Theme and Plot Annotations for 'ggplot2* (2019).

20.Wickham, H. *et al.*. Package 'readr'. (2015).

21.Livingston, E. & Bucher, K. Coronavirus disease 2019 (COVID-19) in Italy. *Jama* **323**, 1335–1335 (2020).

22.Thompson, C. N. *et al.*. COVID-19 Outbreak—New York City, February 29–June 1, 2020. *Morbidity and Mortality Weekly Report* **69**, 1725 (2020).

23.Brynildsrud, O. COVID-19 prevalence estimation by random sampling in population-optimal sample pooling under varying assumptions about true prevalence. *BMC medical research methodology* **20**, 1–8 (2020).

24.Middelburg, R. A. & Rosendaal, F. R. COVID-19: How to make between-country comparisons. *International Journal of Infectious Diseases* **96**, 477–481 (2020).

25.Henning, K. J. What is syndromic surveillance?. *Morbidity and mortality weekly report* 7–11 (2004).

26.Jia, P. & Yang, S. China needs a national intelligent syndromic surveillance system. *Nature medicine* **26**, 990–990 (2020).

27.Zhou, J., Xue, Z., Du, Z., Melese, T. & Boyer, P. D. Relationship of tightly bound ADP and ATP to control and catalysis by chloroplast ATP synthase. *Biochemistry* **27**, 5129–5135 (1988).

28.Boyer, P. D. Energy Life, and ATP (Nobel Lecture). *Angewandte Chemie International Edition* **37**, 2296–2307 (1998).

29.Zinszer, K., Morrison, K., Verma, A. & Brownstein, J. S. Spatial Determinants of Ebola Virus Disease Risk for the West African Epidemic.. *PLoS Curr* **9**, (2017).

30.Rogers, L. C. G., Satchell, S. E. & Yoon, Y. Estimating the volatility of stock prices: a comparison of methods that use high and low prices. *Applied Financial Economics* **4**, 241–247 (1994).

31.R: The R Project for Statistical Computing.

32.Mailund, T. Importing Data: readr. in *R Data Science Quick Reference* 5–31 (Apress, 2019). doi:10.1007/978-1-4842-4894-$2_2$.

33.Wickham, H. & Bryan, J. readxl: Read excel files. *R package version* **1**, (2019).

34.Wilke, C. O. cowplot: streamlined plot theme and plot annotations for 'ggplot2'. *R package version 0.9* **4**, (2019).

35.Schwert, G. W. Stock volatility in the new millennium: how wacky is Nasdaq?. *Journal of Monetary Economics* **49**, 3–26 (2002).

36.Bouzillé, G. *et al.*. Leveraging hospital big data to monitor flu epidemics. *Computer Methods and Programs in Biomedicine* **154**, 153–160 (2018).