ARTICLE OPEN (Check for updates Singapore COVID-19 data cross-validation by the Gaidai reliability method

Oleg Gaidai¹[™], Vladimir Yakimov²[™], Jiayao Sun³ and Eric-Jan van Loon⁴

Novel coronavirus infection (COVID-19) has exserted certain burden on global public health, spreading around the world with reportedly low mortality and morbidity. This study advocates novel bio and health system reliability approach, especially suitable for multi-regional environmental and health systems. Advocated spatiotemporal method has been cross-validated, versus well established bivariate Weibull method, based on available raw clinical dataset. The purpose of this study was to assess risks of excessive coronavirus death rates, that may occur within any given time horizon, and in any region or district of interest. This study aims at benchmarking of the novel Gaidai bio-reliability method, allowing accurate assessment of national public health system risks, for the years to come. Novel bio-system reliability approach is particularly suitable for multi-regional environmental and health systems, monitored for a sufficiently representative period of time. In case when underlying bio-system is stationary, or the underlying trend is known, long-term future death rate risk assessment can be done, and confidence intervals can be generated. Advocated methodology may to be useful for a wide variety of public health applications, thus, it is not limited to the example, considered here.

npj Viruses (2023)1:9; https://doi.org/10.1038/s44298-023-00006-0

INTRODUCTION

Statistical characteristics of COVID-19 (SARS-CoV-2) and other comparable recent influenza outbreaks have been receiving substantial research interest in recent years¹⁻³. Environmental effects on biological systems typically follow cyclical patterns. For environmental effects see ref.⁴; for meteorological parameters see ref.⁵; for heat stress and thermal perception see ref.⁶. In general, determining actual biological system's reliability factors to assess future epidemic outbreak risks, is fairly challenging, given a variety of epidemic and environmental factors. In principle, direct MC (Monte Carlo) simulations or a sufficient number of raw clinical observations might be sufficient to evaluate reliability of a complex biological system. However, COVID-19's clinical observational data are limited to years 2020-2022. In order to address the challenge of having too limited underlying clinical dataset, the authors have developed Gaidai reliability approach, suitable for biological and health systems, when risks of near future epidemic outbreaks are of interest. COVID-19 outbreaks in Singapore were the primary focus of this study, which focused on crosscorrelations between various health data from the same climatic zone. Singapore has been chosen due its extensive national health surveillance, and its publicly accessible raw clinical data⁷. Engineering and medical research both make extensive use of statistical lifetime data modeling by EVT (extreme value theory)⁸. In ref.⁹ EVT has been utilized by authors to forecast H1N1 (swine flu) epidemiological risks. For spatial lag and error models, along with regression techniques, see ref.¹⁰. In the current study an epidemic outbreak is defined as unexpected random event, that might occur at any time and in any administrative region of a particular national health care system. Spatiotemporal aspect of epidemiological risk has been therefore taken into account. Nondimensional parameter λ has been introduced to unify various national regions with different epidemiological backgrounds into one multi-dimensional bio-dynamic system.

Singapore's COVID-19 raw clinical data has been retrieved from a public source⁷. National public health system under investigation has been modeled as MDOF (multi-degree-of-freedom) dynamic biosystem, with strongly correlated administrative components (spatial dimensions). The goal of this study was assessment of future epidemiological outbreaks risks, hence authors only considered daily reported patient numbers, and not symptoms. The map of Singapore represents specific clinical recorded instances.

Based on quasi-stationarity assumption, this study assumed that, despite seasonal fluctuations, the underlying epidemiological process would be statistically representative throughout two consecutive observational years, 2020–2022. In case of underlying trend is of interest, it should be identified first, and epidemiological thresholds should be made variable with time. In the latter case, Gaidai-Yakimov method can be applied even to nonstationary bio-systems.

METHOD

MDOF dynamic system is represented here by a collection of its critical/key components/dimensions, combined into biosystem's representative vector(X(t), Y(t), Z(t), ...), consisting of biosystem's key components $X(t), Y(t), Z(t), \dots$ that has been measured/observed over sufficiently long (representative) clinical period (0,T). Biosystem component's global maxima being $Y_T^{\max} = \max_{0 \le t \le T} Y(t),$ $X_T^{\max} = \max_{0 \le t \le T} X(t),$ denoted as $Z_T^{\max} = \max_{0 \le t \le T} Z(t), \dots$. By sufficiently long clinical/observational duration T authors primarily mean long enough observational duration T with respect to the dynamic bio-system relaxation and auto-correlation time scales. Let X_1, \ldots, X_{N_X} be temporally consequent biosystem component X = X(t) local maxima occurring at discrete temporally non-decreasing time-instants

¹Shanghai Ocean University, Shanghai, China. ²Jiangsu University of Science and Technology, Zhenjiang, China. ³Central Marine Research and Design Institute, Saint Petersburg, Russia. ⁴Ghent University, Ghent, Belgium. ^{Sem}email: o_gaidai@just.edu.cn; YakimovVV@cniimf.ru

 $t_{N_x}^{\chi} < \ldots < t_{N_x}^{\chi}$ within clinical observational period(0, T). Identical definitions can be given for other MDOF bio-system's key components $Y(t), Z(t), \ldots$ namely $Y_1, \ldots, Y_{N_Y}; Z_1, \ldots, Z_{N_Z}$ and so on. For simplicity, all biosystem components, and hence their local maxima have been assumed to be positive. Hence:

$$P = {}^{(\eta_{\chi},\eta_{\gamma},\eta_{z},\dots)} p_{\chi_{T}^{\max},\gamma_{T}^{\max},Z_{T}^{\max},\dots} (x_{T}^{\max},y_{T}^{\max},z_{T}^{\max},\dots) dx_{T}^{\max}dy_{T}^{\max}dz_{T}^{\max}\dots$$
(1)

representing bio-system's survival probability *P*, given in terms of joint PDF (probability density function) *p*. Due to biosystem's high dimensionality, it is not practical to assess $p_{X_T^{max},Y_T^{max},Z_T^{max},\dots}$ directly. When either of key component X(t) exceeds $\eta_{X,r}$ or Y(t) exceeds $\eta_{Y,r}$ or Z(t) exceeds $\eta_{Z,r}$ etc., biosystem is viewed as having instantly failed or entered in a state of hazard. Fixed hazard/failure levels η_X , η_Y , $\eta_{Z,r}$... being individually set for each 1D (1-dimensional) biosystem's component. The latter target biosystem survival probability *P* is needed to assess biosystem's expected lifetime. Bio-system's 1D key components *X*, *Y*, *Z*, ... being now re-scaled as well as non-dimensionalized:

$$X \to \frac{X}{\eta_X}, Y \to \frac{Y}{\eta_Y}, Z \to \frac{Z}{\eta_Z}, \dots$$
 (2)

making all bio-system key components non-dimensional, having identical failure/hazard limits, equal to 1. Synthetic temporally non-decreasing vector being now created by merging/coalescing biosystem component's local maxima into 1D combined system vector $\vec{R} = (R_1, R_2, ..., R_N)$ coherent with corresponding combined temporal vector $t_1 \leq ... \leq t_N$, $N \leq N_X + N_Y + N_Z + ...$. Each biosystem's key component local maxima, constituting vector R_j being actually observed within biosystem temporal record, occurring within either X(t) or Y(t), or Z(t) or other biosystem's components. Constructed synthetic \vec{R} -vector has 0 data loss.

Now temporally non-decreasing synthetic vector \vec{R} , along with its corresponding component's occurrence time instants $t_1 \leq \ldots \leq t_N$, have been now fully introduced^{11–13}.

RESULTS

This section utilizes advocated approach to a bivariate random bio-process Z(t) = (X(t), Y(t)) to demonstrate its efficiency. Patients with COVID who have been diagnosed and daily records are included in this approach, X(t), Y(t), being monitored synchronously over a certain observational time span (0, T). It being assumed for simplicity that samples $(X_1, Y_1), \ldots, (X_N, Y_N)$ within observational time period (0,T) were collected at N equidistant discrete time instants $t_1, \ldots, t_N^{11,12,14,15}$, yielding bivariate joint CDF $P(\xi,\eta) := \operatorname{Prob} (\hat{X}_N \leq \xi, \hat{Y}_N \leq \eta)$ of the 2D vector (\hat{X}_N, \hat{Y}_N) , with components $\hat{X}_N = \max\{X_j : j = 1, ..., N\}$, and $\hat{Y}_N = \max\{Y_{i} \mid j = 1, ..., N\}$. In doing so, it serves as an example of a dynamic two-dimensional (2D) system^{12,13,16}. Using one-dimensional extreme response values with return times and probabilities, critical thresholds were found p. Scaling has been done to combine both time series X, Y in accordance with Eq. (2), resulting in each of the two bio-system components having failure/hazard unitary limit equal to 1. Then, by maintaining them in temporal non-decreasing order, all biosystem components local maxima from each measured system component time-series have been combined into one sinale time-series $\vec{R} = (\max\{X_1, Y_1\}, \dots, \max\{X_N, Y_N\}).$

Synthetic environmental example

The authors selected synthetic example, where exact analytical solution is known in advance. The latter made it possible to crossvalidate advocated reliability method versus well established



Fig. 1 Raw synthetic timeseries along with corresponding \vec{R} vector¹³. Blue and red squares - two original system components X and Y; green squares - synthetic merged vector R.

bivariate Weibull method. Note that Gaidai method can tackle high-dimensional systems, while bivariate Weibull method is suitable only for 2D (2-dimensional) systems. The latter is a distinctive advantage of Gaidai method.

Wind speed 3.65-day maxima process X(t) has been modeled within time period [0, T], based on stationary underlying Gaussian stochastic process U(t), having zero mean value and standard deviation equal to 1. It was assumed for simplicity that U(t) mean zero up-crossing rate equals $v_{11}^+(0) = 10^3/T$, with return period T = 1 year^{14,15,17–19}. As a result, wind speed maxima process X(t)will have $365/3.65 = 10^2$ data points annually, with total data record containing 10⁴ data points, which being equivalent to 100 years. Underlying wind speed process U(t) has 3.65 days maxima analytical CDF (cumulative density function) $F_x^{3d}(x) =$ $\exp\left\{-q\exp\left(-\frac{x^2}{2}\right)\right\}$ corresponding to the 3 days wind speed maxima process $X^{3d}(t)$. Gumbel-Haugaard, Frank, and Clayton are three Archimedean copulas that are often used. The Gumbel-Haugaard copula G(u, v) dependence structure being taken into account initially, modeling cross-correlation between the marginal peak wind speed variables $X^{3d}(t)$ and symmetrically distributed cross-correlated process $Y^{3d}(t)$:

$$G(u, v) = \exp\left\{-[(-\log u)^m + (-\log v)^m]^{\frac{1}{m}}\right\}$$
(3)

with $X^{3d}(t)$ and $Y^{3d}(t)$ having correlation coefficient R_{corr} of 0.5, and parameter $m = 1/\sqrt{1 - R_{corr}}$ being connected to correlation coefficient R_{corr} . Since stationary random Gaussian processes underlying both $X^{3d} = X(t)$, and $Y^{3d} = Y(t)$, the Gumbel-Haugaard copula is easily adaptable, hence bivariate Weibull method prediction agrees well with both analytical solution x = 6, as well as with Gaidai prediction. Exact bivariate CDF reads as:

$$H^{3d}(x,y) = \exp\left\{-\left[q\exp\left(-m\frac{x^2}{2}\right) + q\exp\left(-m\frac{y^2}{2}\right)\right]^{\frac{1}{m}}\right\}$$
(4)

Figure 1 presents simulated (synthetic) time series, coalesced into 1D system \vec{R} vector. Bivariate Weibull contour, with target probability level 2D contour, containing selected bivariate testpoint $(X^{3d}, Y^{3d}) = (6, 5.2)$ agreed well with both analytical and Gaidai method's prediction R = 6, as expected, since underlying stochastic process was rather simple. Second, the equivalent

Clayton copula was used in place of Gumbel-Haugaard copula C(u, v), with asymmetric Archimedean copula:

$$C(u,v) = \max\left\{ [u^{-m} + v^{-m}]^{-\frac{1}{m}}, 0 \right\}$$
(5)

Clayton copula being more challenging for bivariate Weibull method to fit, since it being not part of the copula library currently implemented are only Asymmetric logistic and Gumbel



Fig. 2 Example of Singapore COVID-19 patient's data, presented as daily observed raw timeseries. Red and blue circles - two original bio-system components (number of recorded and dead patients); green squares - synthetic merged vector R.

logistic copulas^{20–26}. Bivariate Weibull method being therefore expected to perform less accurately than Gaidai method in this case.

For specific numerical example mentioned above, it was found that, on average, Gaidai method performed 15–20% more accurately than bivariate Weibull technique. In the case of raw measured non-Gaussian, cross-correlated by non-Archimedean copulas data, an advantage of Gaidai method would be more pronounced. Last but not least, bivariate Weibull clearly required more processing time than Gaidai approach for any given bivariate failure/hazard limit since it performs 2D surface interpolation. Gaidai method has produced 95% CI (confidence interval), while bivariate Weibull method did not have such ability.

Method validation

Figure 2 presents an example of Singapore COVID-19 raw clinical death rate data, recorded during the years 2020–2022, presented as observed timeseries.

Figure 3 presents bivariate Weibull bivariate contours for Singapore COVID-19 death rate data⁷. As seen from Fig. 3, there is an intrinsic inaccuracy, owing to the specific copula choice within bivariate Weibull fit to the raw measured dataset. See for more information on the bivariate Weibull technique^{19,20}. Bivariate failure/hazard test-point (*X*, *Y*) = (44, 000, 65) has been selected for comparison between two methods (Gaidai and bivariate Weibull), as this bivariate test-point lies on the $p=10^{-1.3}$ contour line, predicted by bivariate Weibull technique. 95% CI produced by Gaidai method included bivariate point, utilized by bivariate Weibull method²⁰⁻²³. High-dimensionality (say, above 2D) of biological and health systems makes it challenging to produce accurate multivariate predictions, based on available relatively limited clinical raw datasets. Hence above-described novel health system reliability approach, has advantages of optimally utilizing



Fig. 3 Gaidai method validation versus bivariate Weibull method. a Bivariate Weibull bivariate contours for Singapore COVID-19 data. b Gaidai prediction star indicates the same bivariate failure level of interest.



Fig. 4 Data pattern analysis. a Singapore COVID-19 data 2nd order SODP plot, indicating unnatural correlation pattern. **b** Correlation between COVID-19 daily recorded new patient and deceased numbers. Circle marks outlier.

clinical measured datasets, while taking into account biosystem's high dimensionality.

The Poincare type plot may be used to analyze intrinsic data structural patterns, for example 2nd order difference plot (SODP) can be used to start with. For consecutive differences, 2nd order SODP may be used to statistically observe raw timeseries data²⁴.

Figure 4a presents 2nd order SODP plot. When employing an entropy-based AI (artificial intelligence) recognition approach, 2nd order SODP plots may be used to spot data patterns and compare them to other similar datasets^{14,15,25–27}. This study did not focus on AI pattern analysis as such, therefore Fig. 4a can be seen as motivating for further research, when underlying raw dataset quality remains an open issue.

Figure 4b demonstrates correlation between the daily number of COVID-19 fatalities and newly daily-registered patients. It is clear from Fig. 4b that raw daily recorded new patient counts contain outliers. Traditional health systems reliability techniques that deal with observed raw timeseries do not have an advantage of dealing with high-dimensional (above 2D) systems, along complex cross-correlation between different biosystem components. The key advantage of Gaidai method being its ability to assess reliability of high-dimensional non-linear dynamic biosystems.

DISCUSSION

Traditional timeseries reliability approaches do not always have advantage of easily handling high-dimensional dynamic systems along with cross-correlations between different key system components. Fundamental advantage of Gaidai method being its ability to examine reliability of high-dimensional dynamic biosystems. In this investigation, synthetic wind speeds were used as validation case, as in this case analytical solution are known. The theoretical rationale of the proposed approach being thoroughly discussed. Although using direct measurement or Monte Carlo simulation to analyze the reliability of dynamic bio-systems is often appealing, it should be noted that the complexity and highdimensionality of dynamic bio-systems require development of novel, accurate, and robust techniques that can handle available raw datasets, while utilizing them optimally.

This study's methodology has already been shown successful when applied to a number of simulation models, but only for onedimensional system components. Overall, quite accurate forecasts have been made. The main goal of this study was to develop a general-purpose, trustworthy, and user-friendly multi-dimensional reliability strategy. Gaidai bio-reliability method was compared to the bivariate Weibull method, using both analytically produced synthetic data and actual raw clinical data. To summarize, suggested methodology may be applied to a wide range of biological and public health studies. Presented national public health example by no means limits potential uses of advocated methodology.

DATA AVAILABILITY

Data will be made available on request from corresponding author.

Received: 27 July 2023; Accepted: 22 October 2023; Published online: 13 December 2023

REFERENCES

- Thomas, M. & Rootzen, H. Real-time prediction of severe influenza epidemics using extreme value statistics. arXiv preprint arXiv:1910.10788 https://doi.org/ 10.48550/arXiv.1910.10788 (2019).
- Chen, J., Lei, X., Zhang, L. & Peng, B. Using extreme value theory approaches to forecast the probability of outbreak of highly pathogenic influenza in Zhejiang, China. *PLoS ONE* **10**, e0118521 (2015).
- Mugglin, A., Cressie, N. & Gemmell, I. Hierarchical statistical modelling of influenza epidemic dynamics in space and time. *Stat. Med.* 21, 2703–2721 (2002).
- Sia, A. et al. The impact of gardening on mental resilience in times of stress: a case study during the COVID-19 pandemic in Singapore. *Urban For. Urban Green.* 68 https://doi.org/10.1016/j.ufug.2021.127448 (2022).
- Pani, S., Lin, N. & RavindraBabu, S. Association of COVID-19 pandemic with meteorological parameters over Singapore. *Sci. Total Environ.* 740 https://doi.org/ 10.1016/j.scitotenv.2020.140112 (2020).
- Lee, J. et al. Heat stress and thermal perception amongst healthcare workers during the COVID-19 pandemic in India and Singapore. *Int. J. Environ. Res. Public Health.* 17, 8100 (2020).
- Singapore COVID-19 data. https://voice.baidu.com/act/newpneumonia/ newpneumonia/?from=osari_aladin_banner&city=%E6%96%B0%E5%8A%A0% E5%9D%A1-%E6%96%B0%E5%8A%A0%E5%9D%A1. Acsessed on Jan 2023.
- Thomas, M. et al. Applications of extreme value theory in public health. *PLoS ONE* 11 https://doi.org/10.1371/journal.pone.0159312 (2016).
- Coburn, B. J., Wagner, B. G. & Blower, S. Modeling influenza epidemics and pandemics: insights into the future of swine flu (H1N1). *BMC Med.* 7, 30 (2009).
- Meliker, J. R. & Sloan, C. D. Spatio-temporal epidemiology: principles and opportunities. *Spat. Spatiotemporal Epidemiol.* 2 https://doi.org/10.1016/ j.sste.2010.10.001 (2011).
- Gaidai, O., Cao, Y. & Loginov, S. Global cardiovascular diseases death rate prediction. Curr. Probl. Cardiol. https://doi.org/10.1016/j.cpcardiol.2023.101622 (2023).
- Gaidai, O., Xing, Y., Balakrishna, R. & Xu, J. Improving extreme offshore wind speed prediction by using deconvolution. *Heliyon* https://doi.org/10.1016/ j.heliyon.2023.e13533 (2023).
- Gaidai, O. & Xing, Y. Prediction of death rates for cardiovascular diseases and cancers. *Cancer Innov.* https://doi.org/10.1002/cai2.47 (2023).

- Gaidai, O., Yan, P., Xing, Y., Xu, J. & Wu, Y. A novel statistical method for long-term coronavirus modelling. *F1000Research* https://orcid.org/0000-0003-0883-48542 (2022).
- 15. Gaidai, O. et al. Novel methods for wind speeds prediction across multiple locations. *Sci. Rep.* **12**, 19614 (2022).
- Gaidai, O., Wang, F. & Yakimov, V. COVID-19 multi-state epidemic forecast in India. Proc. Indian Natl Sci. Acad. https://doi.org/10.1007/s43538-022-00147-5 (2023).
- Gaidai, O., Xing, Y. & Xu, X. COVID-19 epidemic forecast in USA East coast by novel reliability approach. *Res. Sq.* https://doi.org/10.21203/rs.3.rs-1573862/v1 (2022).
- Gaidai, O. et al. Cargo vessel coupled deck panel stresses reliability study. Ocean Eng. https://doi.org/10.1016/j.oceaneng.2022.113318 (2022).
- Gaidai, O. & Xing, Y. A novel multi regional reliability method for COVID-19 death forecast. *Eng. Sci.* https://doi.org/10.30919/es8d799 (2022).
- Gaidai, O. & Xing, Y. A novel bio-system reliability approach for multi-state COVID-19 epidemic forecast. *Eng. Sci.* https://doi.org/10.30919/es8d797 (2022).
- Gaidai, O., Yan, P. & Xing, Y. Future world cancer death rate prediction. *Sci. Rep.* 13 https://doi.org/10.1038/s41598-023-27547-x (2023).
- Gaidai, O., Xu, J., Hu, Q., Xing, Y. & Zhang, F. Offshore tethered platform springing response statistics. *Sci. Rep.* **12** www.nature.com/articles/s41598-022-25806-x (2022).
- Gaidai, O., Xing, Y. & Xu, X. Novel methods for coupled prediction of extreme wind speeds and wave heights. *Sci. Rep.* https://doi.org/10.1038/s41598-023-28136-8 (2023).
- Yayık, A., Kutlu, Y. & Altan, G. Regularized HessELM and Inclined Entropy Measurement for Congestive Heart Failure Prediction (Cornell University, 2019).
- Gaidai, O., Cao, Y., Xing, Y. & Wang, J. Piezoelectric energy harvester response statistics. *Micromachines* 14, 271 (2023).
- Gaidai, O., Yan, P. & Xing, Y. A novel method for prediction of extreme wind speeds across parts of Southern Norway. *Front. Environ. Sci.* https://doi.org/ 10.3389/fenvs.2022.997216 (2022).
- Gaidai, O., Fu, S. & Xing, Y. Novel reliability method for multidimensional nonlinear dynamic systems. *Mar. Struct.* 86 https://doi.org/10.1016/ j.marstruc.2022.103278 (2022).

AUTHOR CONTRIBUTIONS

O.G.—conceptualization; V.Y.—writing; J.S.—visualization; E.v.L.—software. All authors contributed equally.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Correspondence and requests for materials should be addressed to Oleg Gaidai or Vladimir Yakimov.

Reprints and permission information is available at http://www.nature.com/ reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



© The Author(s) 2023