Modeling for Explainability: Ethical Decision-Making in Automated Resource Allocation

Christina Cociancig[1], Christoph Lüth[2], Rolf Drechsler[3]

[1,2,3] University of Bremen; German Research Center for Artificial Intelligence (DFKI GmbH), Germany
[1] chrcoc@uni-bremen.de
[2] christoph.lueth@dfki.de
[3] drechsler@uni-bremen.de

Abstract. Decisions delegated to artificial intelligence face an alignment problem: humans expect the algorithm to make fast and well-informed decisions aligning with human morals. In the design and engineering process of algorithms, ethical principles enter the black box explicitly and implicitly as functional or non-functional properties, much to the detriment of explainability and transparency. Previous work has established surrogate modeling to promote explainability and transparency of the decision-making process. We extend on this, model in lower complexity decision trees and as labeled transition systems, which is a method inherent to bisimulation theory, as well as evaluate on synthetic data with a rule-based algorithm. As a case study, we analyze the triage processes in German and Austrian hospitals during the COVID-19 pandemic, based on official guidelines that regulate the allocation of intensive care unit beds. We discovered that the decision processes are similar, however, the systems do not behave in the same manner. The diverging behavior equates to a discrepant ratio of patients treated in intensive care in contrast to the general ward. Our insight leads us to the conclusion that our approach ensures ethical decision-making in healthcare and should be considered due to its explainability and transparency.

Keywords: explainability; transparency; automated decision-making; surrogate modeling.

1 Introduction

Machine learning and artificial intelligence (AI) in general can support virtually any human decision. Whether we assume the decisions made or intervene to revise it, with data supply and automation, we entrust a mostly black box with coming to the best conclusion it possibly can. Often, these decisions we delegate require fast conclusions and a high degree of expert knowledge. While algorithms can fulfill these criteria of decision-making, a human decision is inherently one that draws upon human morals. That we demand of algorithms to emulate this, gave rise to a new field of research that is located at the intercept of computer science and philosophy: machine ethics [1].

Machine ethics, also referred to as AI ethics when defined in a narrower sense, became an increasingly vital factor in AI research, because the decisions we now automate, have direct implications on human lives. Machine ethics places an emphasis on the establishment of values that should steer the development and deployment of artificial intelligences in the form of guidelines for “ethical AI” [2]. Ethicists agree with the pressing issue of ethical algorithmic decision-making by advocating particularly for transparency [3] and explainability [4] of the decisions produced by the black box that a machine learning algorithm, or even more so a deep learning algorithm, can represent.

Related work in algorithmic explainability and transparency put forward various approaches, including but not limited to, surrogate modeling and formal verification. Previous research in the

area of surrogate modeling advanced to complex cases of decision tree modeling of a neural net, with a well-founded result of reduced complexity, high fidelity, and comprehensibility [4]. Even though it has not yet been done in full terms, approximate-bisimulation has been employed to model (dynamic) neural networks and their behavior in terms of their input and output [5]. These approaches add to the extensive list of measures to analyze the decision-making process with the intention of optimizing for ethical decision-making of the system. However, they fail to consider that some explainability and transparency is better than none, especially for use cases that involve critical decisions in healthcare.

AI represents an evolution of informed decision-making in the medical field [6]. In the clinical environment, well informed decisions must be made fast. Not only in Germany this potential has been identified, and discussions to implement decision-making software are well under way or already implemented. SmED (short for "structured initial medical assessment in Germany") is an algorithm that assists medical on-call services to decide where a patient's healthcare needs can be addressed best: a general practitioner or an emergency clinic [7]. While both are not yet applied in the clinical context, OPTINOFA (short for "optimization of emergency care through a structured initial assessment using intelligent assistance services") aims to provide an algorithmic assessment of the urgency of treatment in clinics [8]. The most striking difference between the softwares: SmED appears to be rule-based and is not openly accessible, OPTINOFA is composed of an AI and will be openly accessible.

In the interest of examining a contemporary decision process in healthcare, as a case study, we compare approaches to the decision-process of triage during the COVID-19 pandemic in two countries: Germany and Austria. The decision processes of triage are based on a practice of resource allocation historically attributed to military medicine, which categorizes patients and commonly prioritizes treatment of patients with a high chance of survival [9]. At the beginning of the pandemic, the allocation of resources, i.e., particularly intensive care beds that can accommodate a ventilator, has been regulated by strict guidelines in Germany [10] and Austria [11]. It is exactly this type of situation in which algorithms are capable to provide humans with relief to make well informed, fast decisions. However, it is of the utmost importance that the decisions provided by machines agree with human ethical values.

With this paper we provide a recommendation for transparent and explainable algorithmic decision-making in healthcare, that complies with the ethical principles of explainability and transparency in form of non-functional properties of the system. We build on related work in the area and propose formal surrogate modeling with decision trees, including associated entropy and information gain values indicating the informative strength of a node within the tree, as well as modeling as labeled transition systems, a method inherent to bisimulation, which provides a comparative analysis of the behavior of two systems [12]. With an algorithmic data evaluation, we support the findings of our analysis numerically, and demonstrate, that drawing on metrics inherent to the models provide reference for a comparative analysis of systems. With this recommendation, we hope to contribute an opportunity for ethical healthcare software to be transparent and explainable for medical professionals and patients alike.

2 Methods

This section outlines the methodology applied to construct decision trees and a bisimulation evaluation of triage processes, as well as a description of the algorithm we deployed to measure effects in ratios. Our hybrid approach of two comparative modeling systems and a test with synthetic data was chosen, because it gives a valuable insight into robust tools that can be

accessed for the purpose of investigating strengths and weaknesses of systems such as the triage decision process. We compared and identified differences in the German and Austrian triage guidelines first by focusing on their underlying ethical principles and subsequently in terms of their functional properties.

Both guidelines are governed by implicitly or explicitly defined ethical principals. The Austrian guideline, "Allocation of intensive care resources due to the Covid-19 pandemic", lists four ethical principles influencing every decision within the triage process: justice, non-maleficence, doing good, and the observation of autonomy of the patient [11]. The definition of each principle is linked to several more, some non-ethical, values that shall be upheld: using resources efficiently, allocating fairly, not endangering the supply system, serving the well-being of each individual patient, respecting guardians of patients, and respecting individual freedom [11]. Although the Austrian guideline does not connect these values and principles to individual decisions, each decision made within the process should be guided by them. The German guideline mentions ethical principles predominantly implicitly by connecting them to decisions in the assessment process, including the needs of the patient for intensive care unit (ICU) treatment and the patients will that are directly reflected in decisions within the process, whereas a prohibition of discrimination due to age, social characteristics, and disabilities, as well as fairness are implicit and not represented as decisions within the process per se [10].

Decision Trees

To investigate the sequence of decisions and the associated informative value of decisions, we manually translated the triage guidelines into their respective decision trees and compared the metrics of entropy and information gain, both inherent to information theory. Each decision outlined in the triage guidelines translates into a decision node of a tree. The end nodes represent the decision for or against ICU treatment of an individual patient.

Entropy is measured in bits and represents the average level of information or uncertainty of possible outcomes of a variable. Given a variable X, with possible outcomes $x_1,\ldots, x_n$, with an associated probability of $\mathrm{P}(x_1),\ldots, \mathrm{P}(x_n)$, the entropy of X is defined by Shannon [13] as:

$$\mathrm{H}(X) = -\sum_{i=1}^{n} \mathrm{P}(x_i) \log \mathrm{P}(x_i) \quad (1)$$

Entropy can be calculated for each node in a decision tree. For a description of the strength of the node, the value of information gain expresses the change in information entropy from one node to the next:

$$IG(T, a) = \mathrm{H}(T) - \mathrm{H}(T|a), \quad (2)$$

where $\mathrm{H}(T|a)$ is the conditional entropy of T given the value of a [14]. Information gain can therefore adopt values between zero and one, a higher information gain is associated with a strong decision node, at which an informative decision is made.

Evaluation

To verify the differences in the performance of the two systems, we evaluated benchmark data flowing through the process modeled as decision trees. To this end, we implemented a rule-based algorithm that sorted and evaluated synthetic data of 100 patients based on the health data required for the German triage decision process.

Although the triage criteria to receive intensive care are different for German and Austrian patients, both guidelines have baselines in common, whose negation can in no circumstance lead to treatment in intensive care. For one, a patient must give consent to receive intensive care. Moreover, though the German guideline explicitly states that a necessity for intensive care must be assessed, the same can be assumed for the Austrian system. German medical personnel are furthermore urged to assess the *prospect of success* of intensive care for a patient as one of the first steps in the triage process, whereas Austria assesses *hopelessness* and *proportionality* of ICU treatment only as a criterion for the abortion of intensive care [11]. For our purpose, we equate the assessment of prospect of success in Germany with the assessment of hopelessness and proportionality in Austria and quantify this criterion with 96%, i.e., the average reported survival rate of COVID-19 in Germany and Austria [15].

Beyond the shared baseline assumptions, the triage systems additionally assess the patients on health criteria. These criteria have determined our algorithmic implementation and data development and have been summed to a health score. The criteria formulated for the health assessment of German patients consist of five points, which quantify scores or represent the presence or absence of a criterion: heightened severity of illness, e.g., acute pulmonary embolism, acute organ failure assessed on the sepsis-related organ failure assessment score (SOFA), a prognostic marker for COVID-19 patients (we assume this marker to be a positive COVID-19 test), comorbidity, e.g., neurological disease, and health status assessed on the clinical frailty scale (CFS) [10]. In Austria, health assessment is done in nine points: chance of survival via SOFA score, comorbidity, presence of cardiac insufficiency or failure, renal insufficiency or failure, presence of immunosuppression, dementia assessed on Activities of Daily Living score (ADL), pulmonary disease, other primary disease, and other relevant criteria [11].

As a first step to data creation, we affirmed the baseline assumptions for ICU treatment and created five health data points for each patient randomly, modeling the German health assessment. We randomly one-hot encoded for the presence or absence of a criterion and appoint scores where applicable, i.e., a SOFA score between 0-24, counts of comorbidities between 0-5, and overall health status of CFS score between 1-9. This encoding resulted in an overall health assessment score sum between 1-40, with a higher score being associated with a more critical condition. For scores under 20, we assumed care at the general ward as sufficient, patients with scores over 20 require intensive care. In the corresponding trees, this decision node is represented as a score of under 50% or over 50%.

As a next step we translated the patient data for the Austrian decision process of nine health assessment points, which involved the addition of more scores. As the survival chance in Austria is also indicated by the SOFA score, we adopted it from the German model patient. We transferred comorbidity counts over zero as the presence of a comorbidity and associate a heightened severity in Germany with a primary disease in Austria, as well as translate the CFS score of five and above as the presence of dementia. The remaining criteria, i.e., cardiac or renal insufficiency, immunosuppression and pulmonary disease were again one-hot encoded with a random distribution. The complete assessment results in an overall summed score between 1-31. We again divided the score in over 50% and under 50%, a score of 15 or lower does not receive intensive care.

Bisimulation

To further investigate the difference in behavior of the two triage processes, we remodeled the decision trees as a bisimulation evaluation. Our notation for this evaluation was adopted from Davide Sangiorgi [12]. Specifically, we explore the states and transitions of the processes modeled as labelled transition systems (LTS), which are formally described with triples, i.e., $(Pr, Act, \rightarrow)$ where Pr is a (non-empty) set, also referred to as the domain or the set of the processes of the LTS, Act is the set of actions or transitions, and $\rightarrow$ denotes the transition relation between processes. Bisimulation is a binary relation on the states of two systems P and Q, if for all μ we have:

$$\begin{aligned} &\text{1) for all } P' \text{ with } P \xrightarrow{\mu} P', \text{ there is } Q' \text{ such that } Q \xrightarrow{\mu} Q \text{ and } P' R\ Q'; \\ &\text{2) for all } Q \text{ with } Q \xrightarrow{\mu} Q', \text{ there is } P' \text{ such that } P \xrightarrow{\mu} P' \text{ and } P'\ R\ Q'. \end{aligned} \tag{3}$$

If the bisimulation is complete, meaning for each process in the system P there is an equivalent process in system Q, the systems are bisimilar, i.e., they behave in the same way. If not all processes in system P can be mapped to an equivalent process in system Q, the systems might have equal inputs and outputs, but internally do not behave the same way.

3 Results

Decision Trees

Both triage decision processes were modeled as decision trees. Figure 1 is a comparison of the German triage system and the Austrian triage system. For reference, we added entropy (H) values for each decision node in the trees. Due to our assumption of the baseline criteria for intensive care treatment as being met, i.e., necessity of treatment (represented as the first decision node with $H = 1$ as is standard for decision trees) and consent of the patient, the respective entropy values do not amount to expressive decision nodes.

However, we were specifically interested in the decision node labeled *prospect of success*, as the location of this decision and the corresponding nodes in the trees is an identifiable difference between the two decision processes. Based on our assumption that the prospect of success of treatment, which is assessed early on in Germany, is equivalent to the criterion of hopelessness and proportionality, which is assessed after the ICU treatment has already commenced for the Austrian patient, the anticipated entropy values correspond to the same entropy of $H = .24$. Again, these values are identical, because we assume a survival chance of 96% for both countries, as it is the reported survival rate of COVID-19 in both countries [15]. The information gain, however, of the prospect of success node amounts to $IG = .76$ in the German system, compared to the Austrian system of $IG = .75$.

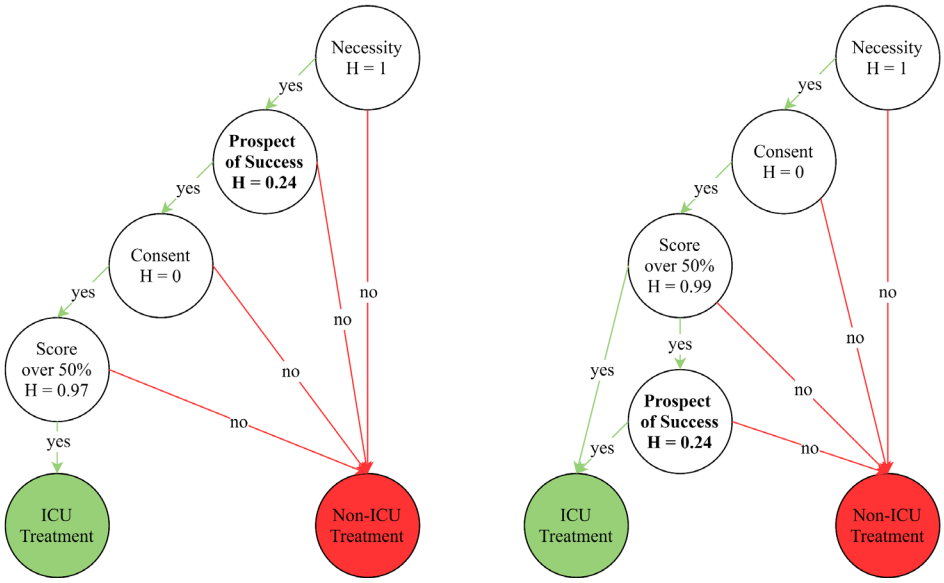

Fig. 1. German (left) and Austrian (right) triage decision tree, entropy value per decision node

Evaluation

The rule-based algorithmic evaluation of our data of 100 patients revealed, that the Austrian system initially treats more patients in ICU, more specifically 56% in comparison to 39% in the German system. This ratio is not representative, however, of the mandatory re-evaluation which is featured in both triage processes and more visible in the labeled transition systems.

Bisimulation

The LTS of triage in Germany and Austria offer valuable cues as to how the decision process is executed and can be described and analyzed formally. Figure 2 compares the LTS side by side and indicates their domain, actions, and transition relations. For the sake of clarity and brevity, we indicated transitions abbreviated, e.g., from R_1 to R_6, instead of "No Necessity assessed" as it would be formally described, we simply indicated "No Necessity".

Not only does the LTS comparison demonstrate that the Austrian triage ultimately has less states until it arrives at a final decision over ICU treatment or no ICU treatment, the formal description of the LTS is a further indicator for the similarity of the systems. As we examine the binary relation between the two systems, we pair processes with the same transition relations:

$$R = \{ (R_1, Q_1), (R_3, Q_2), (R_4, Q_3), (R_6, Q_5) \}. \quad (4)$$

Given that not all states in the German system have equals in the Austrian system, the bisimilarity of the systems cannot be proven and therefore indicate that the systems do not behave in the same way.

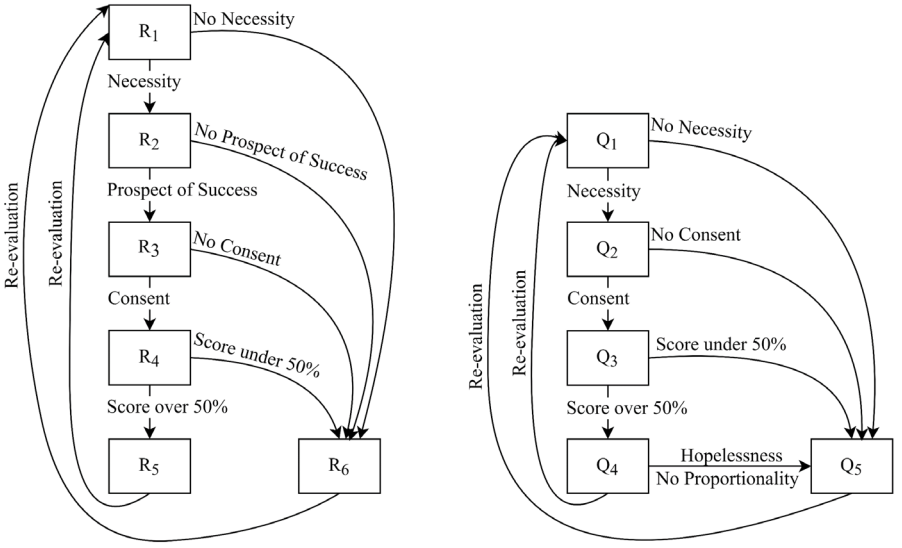

Fig. 2. LTS of triage in Germany (left) and Austria (right)

4 Discussion

The importance of interpretable models, that promote transparency and explainability cannot be emphasized too much. Our method of surrogate modeling as decision trees and labeled transition systems, as well as evaluating on synthetic data, has not only enabled us to identify the different approaches between the triage processes in Germany and Austria, but also given us the possibility to explain any given final decision by traversing through the models. Therefore, by modeling in comparison, we achieve transparent and explainable decision-making as it is promoted by machine ethics [1,3].

At first sight, the systems seem to align in their design, as they assess patients on similar criteria. With the data evaluated based on the decision trees, however, we see that the smallest differences in the processes have a substantial effect on the number of patients treated and thus possibly on the number of lives saved. The decision trees have provided a comprehensible model to enable further evaluation, as we expected and was touched on in [4]. The most striking observation to emerge from the evaluation of information gain, is the difference between the values for the *prospect of success* decision node. Even though the numeric difference between the German and Austrian node is marginally small, it indicates that the informational value of the node is in fact different, and this difference is due to its position in the decision process, i.e., the rank of the decision.

Even though a bisimulation of neural networks has not yet been successful [5], we were able to considerably benefit from modeling the rule-based triage systems as labeled transition systems to reproduce the finding of differences from the decision trees. As the binary relation is incomplete, the evaluation confirms that the decision processes are not equal, and furthermore, that this inequality can be attributed to an inequality in the patient's health assessment and the order of the decisions made in both countries. The model borrowed from bisimulation has

provided us with the ability to formally describe with states and transition relations are represented in both decision processes.

Although we gained considerable insight from modeling the processes, the link to ethical principles, especially to the overarching principles of explainability and transparency of (algorithmic) decisions, were made by us. Neither the German nor Austrian guideline mentions these principles as essential to decisions in triage. Furthermore, despite the insight, we are not qualified to make statements regarding fairness of the decision process, even though the German triage guideline acknowledges this ethical value [10]. Synthetic data is simply not suitable to evaluate this metric on.

5 Conclusion

We have outlined a comparative method of evaluating decision-making by modeling the decision processes in form of labeled transition systems, as well as decision trees and the corresponding metrics of entropy and information gain, to promote the transparency and explainability of decisions within the triage process. For our case study of triage processes in Germany and Austria, we found that differences in the non-functional properties the decision process adheres to, i.e., in broader terms the ethical implications of the triage system, has consequences for the behavior of the system and these consequences can be measured. In the context of triage, the measurable difference of courses of action ultimately equates to lives saved or lost.

For decision-making in triage, we conclude that a formal modelling allows for the analysis and precise comparison of systems. This can be applied to systems of different countries, or two competing systems within one country. Our findings assert that a careful engineering of the decision process, whether with implicitly or explicitly translating ethical principles into decisions, can lead to more efficient decision-making. Besides efficiency, with modeling of the design we access insight about comparability and weaknesses of systems, and measurable metrics can lead to a better understanding of the outcome of specific decisions. In combination, formal modelling, i.e., decision trees and bisimulation, lead to transparency and explainability of the decisions made.

Our method could advance many other decision processes conducted by AI or machine learning in healthcare. Decision-making softwares, whether they include an AI or rule-based algorithm similar to what we employed and SmED appears to be, behave according to underlying ethical principles. As we have concluded from the triage guidelines in our use case, the link between these principles and the properties of the system they are embedded into, are not always functional, i.e., direct and apparent. Yet, the insight gained from the models and their corresponding metrics provides an excellent resource to ensure ethical decision-making. Future work evaluating its models algorithmically on benchmark data should, however, aim to collect or obtain organic data, which was beyond of the scope of this research.

References

1. Anderson, M., Anderson, S. L.: Machine ethics: Creating an ethical intelligent agent. AI magazine 28(4), 15-26 (2007). doi.org/10.1609/aimag.v28i4.2065.
2. Jobin, A., Ienca, M., Vayena, E.: The global landscape of AI ethics guidelines. Nature Machine Intelligence 1, 389–399 (2019). https://doi.org/10.1038/s42256-019-0088-2.
3. Garcia-Gasulla, D., Cortés, A., Alvarez-Napagao, S., Cortés, U.: Signs for Ethical AI: A Route Towards Transparency. arXiv:2009.13871. (2020).

4. Schaaf, N., Huber, M.F., Maucher, J.: Enhancing Decision Tree based Interpretation of Deep Neural Networks through L1-Orthogonal Regularization. arXiv:1904.05394. (2019).
5. Donnarumma, F., Aniello, M., Prevete, R.: Dynamic network functional comparison via approximate-bisimulation. Control and Cybernetics 44(1): 99-127 (2015).
6. Jones, L.D., Golan, D., Hanna, S.A., Ramachandran, M.: Artificial intelligence, machine learning and the evolution of healthcare: A bright future or cause for concern? Bone & Joint Research. 7, 223–225 (2018).
7. Graf von Stillfried, D., Czihal, T., Meer, A.: Sachstandsbericht: Strukturierte medizinische Ersteinschätzung in Deutschland (SmED). Notfall Rettungsmed. 22, 578–588 (2019). https://doi.org/10.1007/s10049-019-0627-8.
8. Abstracts zu Vorträgen und Postern der 14. Jahrestagung der Deutschen Gesellschaft Interdisziplinäre Notfall- und Akutmedizin: 14.–16. November 2019, Bremen. Notfall Rettungsmed. 22, 1–17 (2019). https://doi.org/10.1007/s10049-019-00645-y.
9. Iserson, K. V., Moskop, J. C. Triage in medicine, Part I: Concept, History, and Types. Annals of Emergency Medicine 49(3): 275-281 (2007).
10. Marckmann, G., Neitzke, G., Schildmann, J., Michalsen, A., Dutzmann, J., Hartog, C., Jöbges, S., Knochel, K., Michels, G., Pin, M., Riessen, R., Rogge, A., Taupitz, J., Janssens, U.: Entscheidungen über die Zuteilung intensivmedizinischer Ressourcen im Kontext der COVID-19-Pandemie: Klinisch-ethische Empfehlungen der DIVI, der DGINA, der DGAI, der DGIIN, der DGNI, der DGP, der DGP und der AEM. Medizinische Klinik – Intensivmedizin und Notfallmedizin 115(6): 477-485 (2020). https://doi.org/10.1007/s00063-020-00708-w.
11. ARGE Ethik ÖGARI, Allokation intensivmedizinischer Ressourcen aus Anlass der Covid-19-Pandemie. Klinisch-ethische Empfehlungen für Beginn, Durchführung und Beendigung von Intensivtherapie bei Covid-19-PatientInnen. Vienna: ÖGARI, http://www.oegari.at/web_files/cms_daten/, last accessed: 2021/08/24.
12. Sangiorgi, D.: Introduction to Bisimulation and Coinduction. Cambridge University Press, Cambridge (2011). https://doi.org/10.1017/CBO9780511777110.
13. Shannon, C.E.: A Mathematical Theory of Communication. The Bell system technical journal 27(3): 379-434 (1948).
14. Kent, J.T.: Information gain and a general measure of correlation. Biometrika 70(1): 163-173 (1983).
15. Dong, E., Du, H., Gardner, L.: An interactive web-based dashboard to track COVID-19 in real time. Lancet Inf Dis. 20(5): 533-534. doi: 10.1016/S1473-3099(20)30120-1, accessed: 2021/03/07.