Review and Commentary on: 'Deploying an Artificial Intelligence System for COVID-19 Testing at the Greek Border' *

Harris Georgiou[1]

[1]University of Piraeus[†] – Data Science Lab[‡]
hgeorgiou@unipi.gr

April 4, 2021

Abstract

This is a short review and commentary on a paper that was recently published, regarding the employment of an Artificial Intelligence system for targeted testing of travellers for SARS-CoV-2 at the border checks during the summer of 2020 in Greece. The core thesis of the paper is that the system offered significant accuracy and risk mitigation mechanisms for this task. However, the methodology and the supportive material presented therein shows several deficiencies and biases that do not properly justify such claim. The main problems with the paper are presented here, with short discussion about how they should be addressed by the authors.

Keywords — SARS-CoV-2, COVID-19, epidemics, screening methods, Greece

1 Introduction

On February 26th, 2021, a paper [3] was published by a group of scientists that have been working close to or in the the advisory group of the Greek government regarding the management and policy planning for the national SARS-CoV-2 epidemic for the last year. In their study, the authors present an Artificial Intelligence (AI) system for targeted testing of travellers for SARS-CoV-2 at the border checks during the summer of 2020 and claim that, using this system operationally for those months, the Greek authorities managed to limit the estimated inflow of infected individuals by almost half compared to randomized testing.

Their AI system is based on adaptive estimation of SARS-CoV-2 prevalence in various categories of travellers, primarily based on country of origin from their PLF data registration. The predictive model itself allegedly compensates

*This report is not supported by nor associated with any research project or funding grant.
[†]http://www.unipi.gr – Address: 80 Karaoli & Dimitriou str., GR-18534 Greece.
[‡]http://datastories.org

for scarcity of available tests, delay of results from the laboratories and the low quality of the publicly available epidemic data for each country.

According to the review described in this commentary, the aforementioned study, particularly the statistical and inference aspect of the experimental protocol, presents several deficiencies and drawbacks, which undermine the validity of the authors' core thesis. More specifically, these deficiencies can be grouped around:

- ill-defined core optimization targets
- ill-defined experimental protocol
- non-reproducibility of the results
- ad-hoc assumptions and errors in the methodology
- additional contradictive information

Each of these items is cited and discussed in detail in the following sections.

2 Detailed comments

2.1 Ill-defined core optimization targets

Beginning from the paper's abstract, the authors state the two main targets of their optimization approach:

▶ (pg.1): *'...(i) limit the importation of new cases and (ii) provide real-time estimates of COVID-19 prevalence to inform border policies.'*

The first target implies that, using their AI-based system, the Greek authorities can significantly reduce the inflow of infectious travellers, including asymptomatic ones, hence indirectly addressing the task of risk mitigation for 'safe' opening of the borders during the summer period. In practice, their approach should be able to prove that the level of true positive detections is very high and, at the same time, the level of false negative detections is very low.

The second target implies that, using their AI-based system, the Greek authorities have prevalence estimations of significantly higher accuracy than if using the readily available epidemic data for each country, aggregated and verified via various sources.

Both these claims are very strong and require sufficiently supportive material, in terms of methodology, data used, experimental protocol, reproducibility of results, as well as the statistical significance of the results themselves.

Regarding the first claim, in the abstract the authors continue by stating several numbers in their results:

▶ (pg.1): *'...identified on average 1.85x as many asymptomatic, infected travelers as random surveillance testing, and up to 2-4x as many during peak travel.'*

These numbers refer to counterfactual analysis and comparison against purely randomized testing. Later on, they further elaborate:

▶ (pg.3): *'EVA seeks to maximize the number of asymptomatic, infected travelers identified.'*

► (pg.15)[1]: *'...Our goal is to maximize the expected total number of infections caught at the border.'*

As it is explained below in section 2.2, this claim is over-estimated and not supported by adequate experimental results. In practice, the authors treat the core problem as a gain-optimization task via a multi-armed bandit approach, optimally allocating resources between exploration and exploitation phases. However, gain maximization in the allocated tests, i.e., the 'hit rate' in detecting infected travellers is *not* equivalent to risk minimization of missing imported infections. The reason is that the authors' approach fails to take into account the cost of missing infected cases in the total inflow. In other words, the proposed system would seem to perform perfectly even if only a single 'successful' test was conducted per day, while missing literally all other infected travellers. The proposed method approaches asymptotically a risk-minimization goal only if the testing rate is very high, i.e., when the proportion of untested travellers really becomes statistically insignificant. This deficiency is perhaps the most severe flaw in the authors' paper and overall approach to the problem.

The authors clearly state this discrepancy in their task formulation later on:

► (pg.15): *'...we must balance two seemingly conflicting objectives. On the one hand, a myopic decision-maker would allocate tests to the (estimated) riskiest passengers (...) On the other hand, a forward-looking decision-maker would want to collect data on x-passengers for every value of x (...) This suggests allocating tests uniformly across feature realizations to develop high-quality surveillance estimates.'*

The grey-listing of countries, one of the decision problems addressed in their paper, is based entirely on accurate prevalence estimations. Therefore, it is expected that expectation maximization regarding these country-specific estimations would be part of the formulation, essentially leading to removal of statistical uncertainty, i.e., maximize the tests. There is no explanation by the authors why this option is not considered and, instead, they try to limit the number of tests for 'exploration' and maximize the number of tests for 'exploitation'. In other words, if the incentive is to characterize the risk category of each country, what is the point of conducting probing tests to high-risk travellers when the daily tests available are too few to ensure a realistic no-entry safety margin?

Furthermore, the authors introduce another, non-systemic constraint as a side-objective in their optimization method:

► (pg.4): *'Clearly, grey-listing all countries would minimize incoming infections, but this would also entail a substantive drop in non-essential travel (...), incurring substantial economic costs.'*

Obviously, since this factor is not quantified and included in the optimization formulation described in their paper, it has nothing to do with neither with the methodology or the experimental results. If the global economic loss from grey-listing countries is important enough to balance against missing imported infections from travellers, it should be quantified and included as a cost, or at least as a quantified constraint in the process. If it is not important enough to be included in the formulation, then it cannot be asserted as a justification of preferring the proposed methodology over another 'safer' policy. In fact, ECDC

[1]Note: The paper is separated in two parts, with page numbering restarting after pg.12 where the first part ends. In this text, page numbering is cited continuously, i.e., up to pg.36 at the end of the paper.

clearly states[2] that travel measures and border checking must not be prioritized over public health:

► *'ECDC does not support prioritising travel measures over the public health activities needed in the community such as systematic testing, isolation of cases and contact tracing and quarantine of their contacts. However, If there is no community transmission in a country (if the cases detected in the past 14 days are all imported, sporadic or are all linked to imported/sporadic cases, and there are no clear signals of further locally acquired transmission), some benefit might be achieved by implementing measures at borders.'*

Essentially, the ECDC text in parenthesis says that open-border policies for international travelling should be implemented only if the prevalence of the virus in the country is very low/sporadic. Otherwise, policies for targeted testing at the borders must not be implemented while the national epidemic status is not under full recession.

2.2 Ill-defined experimental protocol

Early on, the authors state their basic claim regarding the quality and usefulness of the global epidemic data that are readily available for each country, as well as the reasons of why not using them instead:

► (pg.1): *However, publicly reported COVID-19 data are imperfect. Different countries follow different reporting protocols and testing strategies. Significant underreporting is known to occur (...) Furthermore, public data generally suffers reporting delays, e.g., due to poor infrastructure.*

Based on this claim, the authors present their AI-based system as providing better, more timely epidemic tracking for incoming travellers. This means that:

- the available data collected at border checks should be statistically more significant, unbiased and current than any other publicly available source;
- the methodology employed to process this data should provide statistically significant and cross-validated results regarding the system's performance metrics.

According to the authors (see pg.15), the average prevalence of SARS-CoV-2 in travellers coming to Greece during the summer of 2020 was in the order of 2 in 1,000 and with large differences between the various countries of origin, hence introducing high imbalance and sparsity in the data (pg.3).

Additionally, the authors state that one of the main constraints of their setup was the scarcity of the available tests. Regarding the average percentage of travellers tested at the border checks:

► (pg.4): *'On average, budget constraints allowed testing 18.3% (±6.1%) of arriving households per day, with a smaller fraction in high-traffic months.'*

However, according to officials[3], during the summer the proportion of random tests conducted on travellers arriving to Greece was even lower, about 10-15% of their total number. This alone essentially invalidates the first optimization target above, i.e., *'limit the importation of new cases'*, unless the

[2]ECDC Questions and answers on COVID-19: Travelling – 4. Why are people not being checked for COVID-19 at the airport when arriving from areas of local or community transmission?

[3]https://www.civilprotection.gr/en/node/6768

authors can prove that even with such a low testing rate they could identify almost every infected traveller in the entire inflow, tested or not, symptomatic or not.

Regarding the data collection and categorization, the authors state that their data correspond to 40 points of entry at the borders (mainly airports and ports) and 193 countries of origin in total, with 38,500 ($\pm$13,590) PLFs processed each day by their system. Even with a testing ratio of 18.3% as stated above, this translates to anywhere between 136,120 and 284,644 travellers daily in total. A more realistic number, according to officials and experience from previous years, should be close to half a million travellers on a daily basis during the peak period, although it is expected that due to the pandemic this was not realized. A very conservative prevalence of 2 per 1,000 translates to 272-569 imported cases per day, probably double during the peak period. Hence, the system should be able to target, test and successfully identify infected travellers at an absolute number very close to these levels. The authors do not provide such absolute numbers in their results, therefore it is very difficult to support that alleged success regarding the first target of their approach.

Moreover, having 193 countries or more (for region-based statistics) data subsets to track daily by processing only a few thousands of PLFs is not sufficiently justified by the authors as adequate for any statistical inference, nor 'better' than using the global epidemic data for each country. In fact, the authors state (see pg.16) that they create 'risky regions' sub-categories, i.e., more than a single label for some (how many?) countries. This further enlarges the pool of types that they have to be sampled with adequately large subset size, in a timely manner (stationarity constraints: 48-hour recent test, 14-days sampling frame), unbiased against other factors (e.g. demographics), and covering all such types. The authors recognize this deficiency in their design, but do not adequately support the claim that somehow their proposed approach overcomes it:

▶ (pg.16): *'...for rare types (...) (e.g., less than 100 arrivals in last 16 days), the variability is quite large. This high variability often renders the estimator unstable/inaccurate, an observation also recognized by prior epidemiological literature.'*

Later on, in the Results section, the authors describe the control for their experiments, i.e., the comparison baseline for evaluating their own methodology:

▶ (pg.5): *'...infected travelers caught by EVA relative to random surveillance testing (i.e., testing an equal number of random arrivals at each point of entry). Since we do not observe the latter quantity, we estimate the counterfactual using inverse propensity weighting (IPW) [39, 40].'*

These statements clarify that: (a) EVA is compared only to pure random 'naive' testing, i.e., not taking into account any epidemic data, and (b) for the most part this comparison is based on counterfactual estimations rather than actual comparative experiments. Although (b) can be justified for some parts of the experimental protocol, when events and time cannot be fully controlled with regard to true epidemic data, (a) is presented without any explanation. Not only EVA is not compared to any other methodology with the same or compatible learning tasks, but no justification is provided regarding why these experiments were not repeated using the global country-specific epidemic data instead of the PLF-based collected in their system.

In addition, Figure 2a in the paper presents two plots of *'estimated num-*

ber of cases caught', but without any units provided for the vertical axis. In other words, no actual number is given regarding how many true positives their system identified during that period. As explained above, if this number is much lower than several hundreds of confirmed cases, then the authors' claim that prioritizing high 'hit' ratio in the testing leads to low inflow of infections is essentially invalidated by simple statistics. Also, Figure 2b refers to testing ratios during 'peak' and 'off-peak' periods at 18.43% and 31.01%, respectively; these numbers are inconsistent to what is described just a few lines earlier in the paper with 18.3% (±6.1%).

Some evidence regarding the true sampling size is provided later on in the description of the methodology:

► (pg.21): *'...we choose thousands of passengers to test in a given day (avg: 5,300; std dev: 998) and receive no feedback on these allocations for 48 hours.'*

In essence, the proposed approach does not even exploits the 38,500 (±13,590) PLFs as stated earlier in the paper but much less, by a factor of more than seven. Even with a testing ratio of 18.3% as stated above referred to this much smaller baseline of 5,300 (±998), this translates to anywhere between 23,508 and 34,415 travellers daily in total, i.e., at least 47-69 imported infections on a daily basis. Again, there is no evidence that the deployed system provided any such number of actual detections per day, as to support the efficiency of the targeted testing policy based on it.

Another contradictive statement is that lower inflow of travellers leads to lower performance (gain) for their system:

► (pg.1): *'...As arrivals dropped, the fraction of arrivals tested increased, thereby reducing the value of "smart" targeting. In the extreme case of testing 100% of arrivals, smart targeting offers no value since both random and targeted testing policies test everyone.'*

This is highly controversial. If the goal of the system is to provide a basis for risk mitigation policies, it is illogical that the system becomes irrelevant as the risk lowers. In fact, the exact opposite is expected for such a decision-support system: more information should lead to better outputs. The real reason for this controversy is, again, the ill-defined optimization target. Since the authors design their system purely on the basis of maximizing the 'hit' ratio while essentially ignoring the real gain of stopping infected travellers, it is inevitable that 100% testing at the border checks renders their system irrelevant. As explained above, the best-performance setup for the proposed system is testing only a single individual successfully with certainty of a 'hit', while ignoring everyone else coming through the border checks.

The authors dedicate an entire section arguing upon the *'(In)-Effectiveness of Commonly Used Public Data'*, saying that:

► (pg.1): *'...these data may not accurately reflect prevalence among asymptomatic travelers (the group of interest for border control policies).'*

They continue stating that their system provides *'the first large-scale dataset for asymptomatic populations across nations'*. However, as described earlier, this is at a scale of 38,500 (±13,590) PLF records per day, for 193 countries plus regional clusters in some cases. In contrast, the publicly available epidemic data for each country contain statistics from several million cases, with few regional data but with daily updates. In terms of statistical significance, it is a very strong claim to make, stating that the first choice is of better quality and reliability than the second one, given the fact that the authors themselves

describe their setup as an optimization task constrained by the scarcity of the tests per day and per entry point. Additionally, the average prevalence of 2 in 1,000 stated by the authors (pg.15) increase the risk of highly biased statistics and uncertainty in estimations when using such small-sized sampling datasets.

In order to justify their choice, the authors provide a series of additional experiments with Gradient Boosting Machine (GBM), Recurrent Neural Networks (RNN) and LASSO regression with two-weeks statistics from the publicly available data for each country. As baseline and ground truth for these models, they employ the corresponding results from their own system:

► (pg.6): *'...We examine the extent to which this data can be used to classify a country as high-risk (more than 0.5% prevalence) or low-risk (less than 0.5% prevalence); (...) We compute the true label for a country at each point in time based on EVA's (real-time) estimates.'* (also detailed in pg.31)

This is also highly controversial: the proposed method is used as the control, instead of a hypothesis to be tested for statistical significance. In other words, no real control is employed in these experiments, only comparative results between the proposed and other methods, asserting the proposed method as the ad-hoc 'correct' output. Furthermore, the proposed method is never actually trained and evaluated upon the exact same data, i.e., the publicly available data for each country without any PLFs for country-specific context. Hence, these experiments cannot validate or invalidate the quality and gain of using this specific data gathering setup, as it is never actually demonstrated that using the publicly available data would produce quite different (worse) estimations. Additionally, the results presented in Figure 3 of the paper are quite fuzzy; GBM is presented as the use case in the plots, RNN and LASSO are referred in a footnote, but no specific numbers (i.e., in a table) are provided for proper comparisons against the proposed method.

The authors also state early on that, via this additional set of experiments, the proposed system is proven as exhibiting higher predictive value in terms of early warning:

► (pg.1): *'...identified atypically high prevalence 9-days earlier than machine learning systems based on publicly reported data.'* (also detailed in pg.32)

They also state that this alleged improvement is attributed to the 'unobserved drivers' that their system discovers in the data:

► (pg.7): *'...These fixed effects collectively model country-specific idiosyncrasies representing aspects of their testing strategies, social distancing protocols and other non-pharmaceutical interventions that are unobserved in the public data, i.e., their coefficients can only be inferred by using EVA's testing results. The improvement imbued by Model 5 suggests that these unobserved drivers are critical to distinguishing high- and low-risk countries.'*

Again, this claim is never proven by comparative experiments and results, using different predictive models but *trained on the same data* and/or a single model trained on different data (global or PLF-specific). The hypothesis is ill-defined and, most importantly, the evidence provided for supporting this claim is statistically invalid.

Specifically for the delays in reported cases, the authors again use their own method results as the control:

► (pg.7): *'...For each country, we use case data to predict its current risk status y_t, i.e., whether its true prevalence (as measured by EVA) exceeds its median true prevalence over the summer.'* (also detailed in pg.32)

As for the previous case, it is invalid to employ the proposed method as the baseline for any similar hypothesis check. If the predictor to be validated is the provider of the output for checking the estimated lag, then this estimation cannot be asserted ad-hoc to be valid and reliable. Instead, other standard methods should be used for this process, including e.g. time-series analysis, auto-regressive models, spectral decomposition, etc.

Next, counterfactual analysis is employed to estimate the effects of grey-listing specific countries and how this affects the data and statistics collected for them along the temporal scale. However, this analysis is heavily based on ad-hoc assumptions that cannot be fully justified, according to the explanation above:

► (pg.7-8): *'...Thus, to quantify the benefit of early grey-listing, we must create counterfactual estimates of both the prevalence and arrival rates had a country not been grey-listed. We form such estimates by fitting GBM (...) we assume that any country that EVA recommended to greylist would also have been grey-listed by a baseline surveillance system, but 9 days later.'*

In terms of signal processing and time series analytics, the GBM is a non-linear forward predictor, not employing nor implying any strict causality between the (training) input and the (desired) output. In other words, the exact same or similar predicted time series can be provided by similar regressors but for time lags quite different than the 9 days. Hence, this is not a valid comparison baseline for estimating the performance of the proposed method. Additionally, in Figure 4 of the paper the vertical axes in the plots have no units, which makes it impossible to actually compare the cases presented therein. Even in the peak season, the authors state that their proposed method *'prevented an additional .12x (±.022) asymptomatic, infected travellers'*, which is a very small number in terms of statistical significance, given the fact that the average prevalence in the travellers is already very low and the daily datasets relatively small in size, as explained earlier. Finally, this 'additional' factor is inherently invalid if not properly formulated against other possible joined probabilities, i.e., conditionals that may already be highly correlated with it, as Bayes theory requires.

Besides the primary decision problem of allocating batches of tests to entry points (see pg.13), the authors define a secondary decision problem as choosing 'color designations' for the inherent risk level for each country of origin (see pg.14). This is not different than what they could do similarly when using the publicly available epidemic data for each country. Therefore, any optimization/learning task associated with this secondary decision problem should also be studied using this alternative pool of data, for comparative results. Nevertheless, the authors do not conduct such comparative analysis and, furthermore, limit the task to only grey-listing suggestions for countries and with informal side-feedback from the Greek COVID-19 taskforce. This means that this particular decision problem is ill-defined, not properly formulated and fuzzy in terms of actual objective evaluation against other options or algorithms.

Additionally, Figure 1 (pg.14) illustrates the main points of entry that were considered in their study. These do not seem to include any land border checkpoint in Epirus or Macedonia, which are also regions with high rates of tourist inflow during the summer from neighbouring countries. In the caption the authors state:

► (pg.14): *'...Selected points of entry (...) Note that information on some points of entry is omitted due to the sensitive nature of the data.'*

Clearly, this information obfuscation is incompatible with a proper scientific publication. If the exact details of the 40 points of entry, one of the main components of the problem formulation, is not clearly documented, then the quality of the collected data and the performance of the system itself cannot be properly assessed.

Another issue of unclear description in the experimental work of the study is comparisons and performance estimations, e.g. using MSE in Figure 2 of the paper (pg.17). Normally, a proper baseline is used as control and additional outputs are compared to it. However, in Figure 2, as well as in the equation above, it is not clear what is compared against what:

► (pg.17): *'...As a simple illustration, consider a (foolish) baseline estimator that estimates every prevalence to be zero. We compute this baseline estimator and the naive estimator for all countries (...) In other words, any potential signal is entirely washed out by noise.'*

In other words, the authors confirm that having such low average prevalence of infected travellers in the inflow essentially makes their estimation process invalid. In order to prove that their proposed approach can somehow 'discover' the elusive information content that is 'washed out by noise', the authors should compare it at least with this 'naive' estimator, not against 'blind' random testing that is based on pure chance. No such evidence is presented in the paper. Additionally, the authors state that:

► (pg.17): *'...Our approach naturally allows information-sharing across types, so that rare types partially borrow data from other similar types to improve their stability.'*

This means that for under-represented types in the daily sample pool, some types (perhaps entire countries?) are essentially merged together for statistical reasons. Clearly, this invalidates the authors' claim that their PLF-based approach is qualitatively better in terms of statistical validity and accuracy compared to using the publicly available epidemic data for the countries.

Regarding the performance baseline used, it is clearly stated again later on:

► (pg.26): *'...we benchmark against a random surveillance testing policy that tests passengers uniformly at random at each port of entry.'*

As already explained, random testing is an invalid baseline for comparison. Instead, the authors should demonstrate the performance of their proposed approach using: (a) the same PLF-based data with different methods/algorithms and (b) their own method with the publicly available epidemic data. The first part would demonstrate the validity and success rate of their methodology against others, while the second part would demonstrate the gains from using PLF-based data instead of other open sources. Nevertheless, neither of these are presented in the paper.

Furthermore, the authors claim again that PLF-based data provided surveillance of better quality:

► (pg.26): *'...As we show in Section 4, such public data offers limited predictive value for identifying high-risk countries, and further carries an information delay of approximately 9 days (...) Due to operational limitations, we do not know the actual number of type k travelers who arrived at entry point e on day t. Hence, we estimate this number as follows (...) Due to no-shows, this number may be less than $T_{ke}(t)$ (...)'*

The first note that should be made here is that the discussion about no-shows, inaccurate $T_{ke}(t)$ and yet another estimation step essentially invalidates

the argument of using PLF-based data as of better quality. It is very clear that this was not true in the available infrastructure, resources and time constraints. The second note is that under no circumstances this experimental protocol can be asserted as valid and reliable when the 'operational limitations' do not permit proper logging of arrivals per entry point. If this is true, then essentially everything related to the sample data used in the study is subject to statistical bias, incompleteness or insignificance with regard to experiments. Furthermore, the authors do not even provide numbers for the type-specific fraction $s_{ke}(t)$ of passengers that actually arrived (pg.27).

Regarding the regressor-based experiments (pg.31) and the clustering (pg.32-33) of countries based on their estimated delay of prevalence in publicly available epidemic data, the authors describe an empirical-only approach of estimating the clustering performance. Specifically, they fail to provide quantitative results on the clustering using some standard metrics, e.g. silhouette. Additionally, they do not explain why they do not employ a standard clustering method as the baseline or at least for comparison purposes, e.g. k-means. Finally, they do not provide comparisons to other more appropriate methods for estimating time offset discrepancies, e.g. cross-correlation, Dynamic Time Warping (DTW), Akaike information criterion (AIC), etc. Based on the results provided, the authors do not sufficiently support their claim that using their system with PLF-based data resulted in better, lag-invariant output than using such regressor-based models that use publicly available epidemic data (no such comparison is presented).

2.3 Non-reproducibility of the results

In the final notes of the first part of their paper, the authors note that neither the data or the full implementation (code) of their proposed methodology is currently available:

► (pg.12): *'...Data availability. The data that support the findings of this study are available from the Ministry of Digital Governance but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission from the Ministry of Digital Governance.'*

Based on this statement, it is clear that all experiments and results presented in the paper, especially everything that relates to PLF, availability of tests per day per entry point, positivity rates, etc, are not publicly available. Hence, these experiments are non-reproducible by anyone else. As a result, since a significant part of the authors' work presents claims based exactly on these context-specific data that differentiates them from the publicly available epidemic data for each country, much of the study cannot be validated either.

Regarding the implementation of their methodology:

► (pg.12): *'...Code Availability. The code for the deployment of the algorithm on a sample dataset is available at (...)*[4]*. The code for the counter-factual analysis is under preparation and will be available at (...)*[5]*.'*

The first repository is a package of R sources, last updated on November 23rd, 2020. It contains four R scripts related to the bandit algorithm, the

[4]https://github.com/vgupta1/EvaTargetedCovid19Testing
[5]https://github.com/vgupta1/EVA_CounterfactualAnalysis

daily run for tests allocation, some feature selection for countries and a few helper functions for data import/export, for a total of four source files (.R) with 699 lines of code including comments and headers. Hence, several items of the methodology and experiments presented in the paper are missing, e.g., the GBM/RNN/LASSO comparative experiments. Also, all data provided in the same package are clearly described as samples, not related to the real operational deployment of the system. The second repository is related to the counterfactual analysis, last updated on February 12th, 2020, and is still empty of any code. Hence, based on these two cited sources, neither implementations are available for review and repetition of the described experiments, even with data different than the PLF-based as described in their paper.

2.4 Ad-hoc assumptions and errors in the methodology

In section Methods the authors provide details about the exact problem formulation, the design of their proposed approach, the optimization/learning targets and the constraints.

In a footnote, the authors clarify that demographic features about age and gender are not considered as important:

► (pg.13): *'Note that we do not include age and gender features since we did not find them to be predictive. This is likely due to the operational constraint that all members of a single household (i.e., spanning different ages and genders) are included in the same PLF and at most one member (per PLF) is tested.'*

The first note that should be made here is that relevance between gender and age with asymptomatic carriers of SARS-CoV-2 is still under debate. Although it seems that these demographic factors may have little statistical correlation to the virus prevalence in the general population, there are other studies that provide such evidence for Italy [2] and South Korea [4]. Hence, these should also be included in the paper, investigated thoroughly via data-driven analysis and documented as statistically relevant or not.

The second, more important note is that the authors clearly state here that one PLF record is used per household. This essentially means that: (a) every family member is being considered equally, i.e., either infected or not, without any further investigation, and (b) every family is treated as a single 'draw' in terms of testing for positive or negative result. Since no scientific evidence has proven that the virus prevalence within a family is all-or-none, these two facts essentially downsample the testing pool by a factor of 3-5 on average, assuming that most travellers coming to Greece are within families of that size. In practice, the authors treat all features about gender, age group and family association more or less as indifferent to the learning task at hand.

In another note on their methodology, the authors claim that employing Bayesian (posterior) estimates addresses this drawback (Eq.2-4, pg.18). However, this is not true, since Bayesian estimators [6] and learning methods [7] can only exploit correlations to improve posteriors, not compensate the information loss in too-small sampling pools. Therefore, while the description of the posterior updates ('Estimation Strategy' algorithm, pg.19) is valid, it does not prove its efficiency in the extremely limited sample data, as described in the paper.

Regarding the identification of 'risky regions' as more precise type identification than plain country-based granularity, the authors state that:

► (pg.20): *'These additional types allow us to exploit intra-country heterogeneity in prevalence to better allocate testing resources. We identify risky regions using the celebrated LASSO procedure (...)'*

There is no clear description on how this is achieved in terms of size of pooled data, i.e., how this learned taxonomy is built upon very limited sample data available on a daily basis, as described above. Furthermore, 'LASSO logistic regression' is described as the selected method to do this. However, the model described in the equation therein for y_i and the 'Adaptive Definition of Types' algorithm next (pg.20) is pure linear, not logistic (see [5] for proper definition). Moreover, in the definition of this linear model no weighting factors are employed for proper regularization between the two factors affecting y_i. As a result, the sparsity sub-vectors δ are heavily affected by the f_i and c_i instances, which in turn can make the Bayesian estimate r irrelevant. These errors raise serious questions about the validity of the description, as well as the performance of the regressor, in the task of identifying these 'risky regions' in the context of the proposed approach.

Regarding the core idea of the methodology, employing 'exploration' and 'exploitation' phases, the authors state:

► (pg.21): *'...one can confirm that Eq.(5) naturally balances the twin goals of exploration (prioritizing types with a wide prior, i.e., large variance) and exploitation (prioritizing types with large, estimated prevalence.'*

As explained above, the authors do not sufficiently justify why targeting for variance minimization is not the sole goal of their approach. In other words, if the general intention is to minimize the number of imported infections in terms of maximum likelihood estimates of prevalence per type (country or region), then risk minimization is inevitably depended on minimizing the uncertainty, i.e., variance, of these estimates. In contrast, the authors' approach 'spends' much of the testing resources in maximizing the 'hit' ratio during the exploitation phase, which is practically useless with such low screening capacity, as the authors themselves recognize. The proposed method samples 1 in 5 at best and asserted as capable of drastically limiting the imported cases from the entire pool, a claim that is not proved in the paper. In addition, having this 48-hour delay in feedback for the testing results, the authors do not provide a post-analysis of their system's output within these two days, in order to quantify its true performance, although this should be a straight-forward and expected result to present in their experimental protocol.

Regarding the pseudo-update for the Gittins index, the authors state:

► (pg.21): *'...Although we do not observe immediate feedback when allocating a test, we can estimate the likely reduction in the variance of our posterior distributions (...) Importantly, pseudo-updates allow the optimistic Gittins indices to dynamically change during the course of allocation assignment within a batch.'*

Although description of the optimality of the Gittins index is provided and cited, pseudo-updates invalidate this evidence: There is no guarantee that these estimations are accurate enough as to not invalidate the optimality of the Gittins index itself. In other words, using erroneous, biased or noisy posterior distributions as input to the (optimal) Gittins index formulation according to the 'Gittins Pseudo-Update for type k' algorithm (pg.22) does not make the entire process optimal too. The authors should provide such proof or cite relevant sources to support this claim.

Subsequently, the authors describe the option of 'prior widening' of variances, a technique that artificially inflates the initial range of 'guess' prior to Bayesian updates. Despite the fact that this alone invalidates the intuitive goal of limiting, not inflating, the uncertainty of prevalence per type (see above), the authors also state:

► (pg.23): *'...We periodically tuned c to ensure that every type with sufficient arrivals was allocated at least 500 tests every 16 days: this is roughly the number of tests required to distinguish a type with 0.5% prevalence from a type with 0.1% prevalence with high probability.'*

This estimation is not explained properly: Based on what priors? What posterior distributions (models)? What does 'with high probability' mean? How does this relate to the number 500? What is the exact estimation method employed to confirm this, e.g., via confidence intervals? These are important items, related to the formulation and correctness of the proposed method, yet they are never explained in the paper.

Regarding the operational applicability of their system, the authors describe the imposed constraints and operational limitations:

► (pg.24): *'...An optimal allocation can be determined by solving a large binary integer program, but solution times can be long – (footnote) Passengers can fill out PLFs up to the day before travel; at the same time, EVA had to decide all testing allocations at the start of the day. As a result, EVA had to output all test allocations within 1 minute of receiving the PLFs for the day to be operationally viable.'*

This description is not realistic. A simple artificial cutoff time could be imposed in the data feed, e.g. use PLFs up to 23:00' instead of 00:00' (midnight), thus providing the same results with only one-hour time slip in the overall scheduling. Regardless of the operational use, the authors could easily implement any other methodology and algorithm for post-analysis and comparison with their proposed approach. For example, if the pure logistics/operational target is to be evaluated, the full binary integer solution could be used with past data as the performance baseline. The same comparative experiments could be implemented for the risk minimization target regarding the high-rate detection of imported infections, comparing the proposed approach to the actual data that became available a few days after the tests or if travellers were later registered as confirmed infections in Greece or after returning to their country. None of these cross-validation results are presented in the paper.

One of the most severe flaws in their methodology is stated by the authors themselves:

► (pg.28): *'...assuming that the probability EVA tests a type k passenger at time t at point of entry e is independent of previous testing results (again, conditional on the arrival process of passengers). Strictly speaking, this assumption does not hold because the bandit allocations at time t depend upon the estimated prevalence, which depend on the infection status of individuals tested in the last 14 days of test results.'*

The authors assert this independence assumption in order to estimate the variances required for the estimation of the total number of positive cases that random surveillance would have caught. This assumption invalidates the value of the entire learned model in their proposed approach, as the 14-day backlog becomes somehow irrelevant of their algorithm's output. Furthermore, the arguments to support such claim, provided therein by the authors, are over-simplified

and without proof. For example, they state that employing a minimum size of *'500 tests every 14 days'* for each type is *'a large number of exploration tests'*, which is not adequately supported, as explained above. This claim is further complicated by another assertion, stating that *'if we test a large number of passengers of a particular type, our estimated prevalence (...) will be very close to the true prevalence'*, which in fact is an argument of using the publicly available epidemic data instead of the (limited) PLF-based data.

Another severe flaw in the authors' methodology is the application of counterfactual analysis to the value of grey-listing. In particular, the authors state that:

► (pg.30): *'...Notice all these infections were prevented by EVA since none of them arrived in Greece (they remained home and did not travel).'*

In essence, the claim is that, since the deployed system suggested grey-listing of specific countries and this resulted to lower inflow from these countries or regions, the system should be given credit for even larger increase in screened infections. Besides being pure speculation, this line of thinking has several logical flaws: (a) the drop in inflow due to grey-listing is based, again, on estimations via regression, not real data; (b) all decrease in inflow is accredited to the system, leaving out any other context or factor, again without any supporting data; (c) the no-shows sample from the grey-listed countries or regions is assumed to be statistically indifferent from what was available prior to grey-listing, i.e., these countries/regions are assumed to be epidemiologically stationary, hence no adaptive learning task should be required in this case.

2.5 Additional contradictive information

Recently, a new study [1] was published with results regarding the prevalence of SARS-CoV-2 in travellers returning to UK during the summer of 2020. According to this study:

'...4,207 travel-related SARS-CoV-2 cases are identified. 51.2% (2155/4207) of 69 cases reported travel to one of three countries; 21.0% (882) Greece, 16.3% (685) Croatia 70 and 14.0% (589) Spain. (...) The highest number of cases and onward contacts were from Greece, which was largely exempt from self-isolation 120 requirements (bar some islands in September at the end of the study period).'

The study continues with clarifications regarding the time period in question, where most of the infectious probably took place outside the UK (in Greece):

'Importations from Greece came at the end of August and continued into September, with the steepest of all curves. No travel restrictions were imposed on Greece during this time period and it was the source of greatest imported SARS-CoV-2 cases during this study period.'

What these results show is that, in fact, prevalence of the virus in the main tourist regions in Greece during the summer of 2020 was severely underreported, unchecked and not limited by the targeted testing policy employed at the border checks. Figure 1 presents plots with the frequency of importations over time for the top 4 most common countries of travel reported by individuals testing positive for SARS-CoV-2 during the study period. The shaded areas represent the period of time when the countries did not have restrictive travel guidance in place. Besides being the country with the highest rate of imported (to UK) confirmed infections after the summer of 2020, Greece is also the country with the largest non-restricted travel period at that time among these top

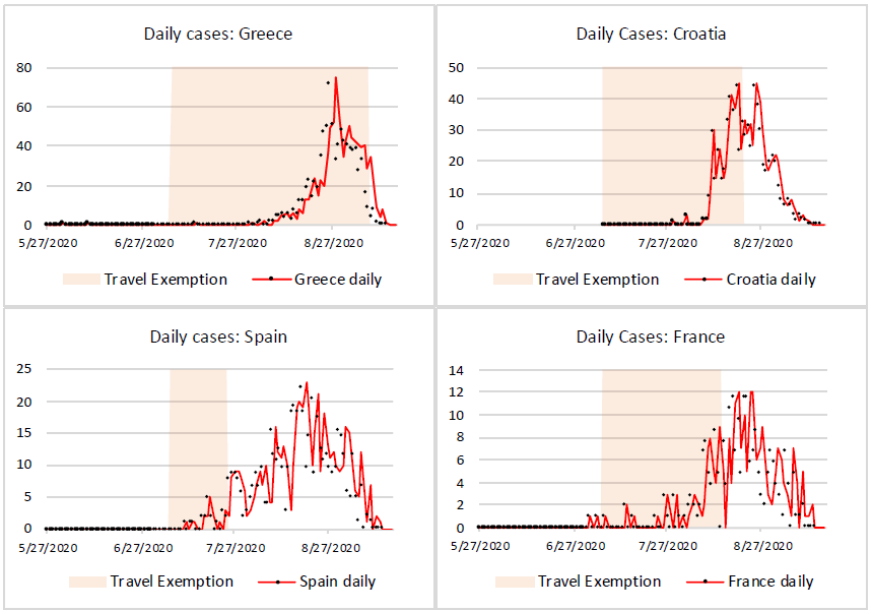

Figure 1: Frequency of importations over time for the top 4 most common countries. The shaded areas represent the period of time when the countries did not have restrictive travel guidance in place. (source: [1])

4 countries.

Even if these infections to tourists are considered outflow, i.e., arrived healthy and infected in Greece, the entire methodology and experimental protocol presented in the authors' paper is put into question: Since for the most part of the summer Greece exhibited a low rate of daily infections, and if the targeted testing policy at the border checks worked as efficiently as claimed, then there should be no subsequent surge of the national epidemic, as it actually did, beginning from the middle/end of September and onto October-November. In other words, if the epidemic in Greece was in full recession during the summer and the employed system screened the infected travellers effectively, this severe second wave of the virus would never occur.

3 Conclusions

In this short review, detailed comments were provided regarding the core thesis and the supporting evidence presented in the study in question. Several problems were identified in the statistical aspects of the methodology, the material and the experimental protocol employed by the authors. More specifically:

1. The core optimization goal is incorrect; maximizing 'hit' ratio in tests conducted does not guarantee risk minimization of missed inflow of infected travellers.

2. The authors never explain what is the true value of the 'exploitation' phase, given the fact that less than 1 in 5 travellers (at best) were screened

and, thus, no strong statistical evidence is provided that nevertheless it was sufficient to detect literally all infected individuals at the border checks.

3. The use of PLF-based data for prevalence estimation, instead of publicly available epidemic data for each country, is never proved as beneficial and of better statistical quality.

4. Moreover, the small size of the PLF-based data and the treatment of each household as one unit (same PLF) essentially invalidate many of the base assumptions for the problem formalization and the learning process in the proposed approach.

5. No comparative experiments are conducted with regard to different algorithms using the same training data, as well as the proposed approach trained with different training data (PLF-based versus publicly available).

6. Very limited or no quantitative metrics were used in several aspects of the performance evaluation; moreover, the proposed system is asserted as the comparison baseline, instead of itself being compared to other realistic baselines.

7. The main comparison in terms of validity and performance of the proposed approach is versus the pure random selection, i.e., not even using the global trends from publicly available epidemic data for each country. Obviously, this cannot be used alone as adequate supporting evidence.

8. No numbers are provided in the results regarding actual detections via the proposed system, rate of arrivals per entry point, final number of types (countries, regions), number of samples per type within the active time frame, etc.

9. No full post-analysis is provided for the look-ahead functionality of the proposed approach, specifically the pseudo-updating and the optimality of the Gittins index after that.

10. No full post-analysis is provided without the operational time constraints, in comparison to more time-consuming but more precise alternative methods, as a baseline for the true validity and performance of the implemented system.

11. The implemented system is published only partially and with no availability to the real data used. Hence, reproducibility of the experiments, the results and the conclusions of the authors' paper is purely speculative or simply impossible.

12. No explanation, nor reference to, is provided by the authors regarding evidence suggesting that travellers returning from Greece exhibited very high virus prevalence. If national epidemic was in recession throughout the summer and the targeted screening at the border checks was as effective as described, the second surge that hit Greece after mid/late September would never happen.

Since the specific publication is not yet submitted for peer review but yet it is of high impact, it is expected that these drawbacks will be addressed successfully by the authors. This commentary provides some hints towards this direction.

References

[1] D. Aggarwal, A.J. Page, U. Schaefer, G.M. Savva, and et.al. An integrated analysis of contact tracing and genomics to assess the efficacy of travel restrictions on sars-cov-2 introduction and transmission in england from june to september, 2020. *medRxiv*, 2021.

[2] I. Arisi and E. Mantuano. Age and gender distribution of covid-19 infected-cases in italian population. *(Preprint)*, 2020.

[3] H. Bastani, K. Drakopoulos, V. Gupta, and et.al. Deploying an artificial intelligence system for covid-19 testing at the greek border. *SSRN*, 2021.

[4] C.-Y. Chan-YoungJung, H. Park, D.-W. Kim, and et.al. Clinical characteristics of asymptomatic patients with covid-19: A nationwide cohort study in south korea. *International Journal of Infectious Diseases*, 99, 2020.

[5] S-M Kim, Y Kim, K Jeong, H Jeong, and J Kim. Logistic lasso regression for the diagnosis of breast cancer using clinical demographic data and the bi-rads lexicon for ultrasonography. *Ultrasonography*, 37:36–42, 2018.

[6] M.R. Spiegel, J. Schiller, and R.A. Srinivasan. *Probability and Statistics (3rd/Ed.)*. McGraw-Hill, 2009.

[7] S. Theodoridis. *Machine Learning: A Bayesian and Optimization Perspective*. Academic Press, 2nd edition, July 2020.