Article

Efficient and targeted COVID-19 border testing via reinforcement learning

Hamsa Bastani[1,7], Kimon Drakopoulos[2,7 ✉], Vishal Gupta[2,7], Ioannis Vlachogiannis[3], Christos Hadjichristodoulou[4], Pagona Lagiou[5], Gkikas Magiorkinis[5], Dimitrios Paraskevis[5] & Sotirios Tsiodras[6]

Throughout the coronavirus disease 2019 (COVID-19) pandemic, countries have relied on a variety of ad hoc border control protocols to allow for non-essential travel while safeguarding public health, from quarantining all travellers to restricting entry from select nations on the basis of population-level epidemiological metrics such as cases, deaths or testing positivity rates[1,2]. Here we report the design and performance of a reinforcement learning system, nicknamed Eva. In the summer of 2020, Eva was deployed across all Greek borders to limit the influx of asymptomatic travellers infected with severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), and to inform border policies through real-time estimates of COVID-19 prevalence. In contrast to country-wide protocols, Eva allocated Greece's limited testing resources on the basis of incoming travellers' demographic information and testing results from previous travellers. By comparing Eva's performance against modelled counterfactual scenarios, we show that Eva identified 1.85 times as many asymptomatic, infected travellers as random surveillance testing, with up to 2–4 times as many during peak travel, and 1.25–1.45 times as many asymptomatic, infected travellers as testing policies that utilize only epidemiological metrics. We demonstrate that this latter benefit arises, at least partially, because population-level epidemiological metrics had limited predictive value for the actual prevalence of SARS-CoV-2 among asymptomatic travellers and exhibited strong country-specific idiosyncrasies in the summer of 2020. Our results raise serious concerns on the effectiveness of country-agnostic internationally proposed border control policies[3] that are based on population-level epidemiological metrics. Instead, our work represents a successful example of the potential of reinforcement learning and real-time data for safeguarding public health.

In the first wave of the pandemic, many countries restricted non-essential travel to mitigate the spread of SARS-CoV-2. The restrictions crippled most tourist economies, with estimated losses of US$1 trillion among European countries and 19 million jobs[3]. As conditions improved from April to July, countries sought to partially lift these restrictions, not only for tourists, but also for the flow of goods and labour.

Different countries adopted different border screening protocols, typically based on the origin country of the traveller. Despite the variety of the protocols, we group those used in early summer 2020 into four broad types: allowing unrestricted travel from designated 'white-list' countries; requiring travellers from designated 'grey-listed' countries to provide proof of a negative test by PCR with reverse transcription before arrival; requiring all travellers from designated 'red-listed' countries to quarantine on arrival; forbidding any non-essential travel from designated 'black-listed' countries.

Most nations employed a combination of all four strategies. However, the choice of which 'colour' to assign to a country differed across nations. For example, as of 1 July 2020, Spain designated the countries specified in ref. [1] as white-listed, whereas Croatia designated these countries as grey-listed or red-listed.
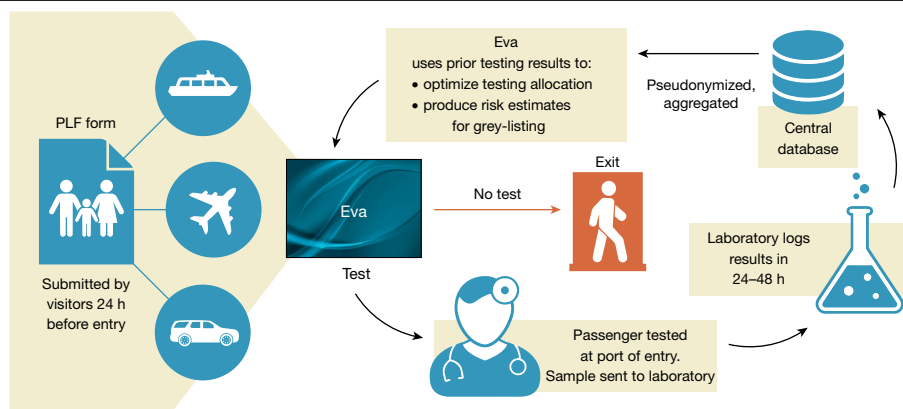
To the best of our knowledge, in all European nations except Greece, the above 'colour designations' were entirely based on population-level epidemiological metrics (for example, see refs. [1,2]) such as cases per capita, deaths per capita and/or positivity rates that were available in the public domain[4–6]. (An exception is the UK, which engaged in small-scale testing at select airports that may have informed their policies.) However, such metrics are imperfect owing to under-reporting[7], symptomatic population biases[8,9] and reporting delays.

These drawbacks motivated our design and nationwide deployment of Eva: the first fully algorithmic, real-time, reinforcement learning

Fig. 1 | A reinforcement learning system for COVID-19 testing (Eva). Arriving passengers submit travel and demographic information 24 h before arrival. On the basis of these data and testing results from previous passengers, Eva selects a subset of passengers to test. Selected passengers self-isolate for 24–48 h while laboratories process samples. Passengers testing positive are then quarantined and contact tracing begins; passengers testing negative resume normal activities. Results are used to update Eva to improve future testing and maintain high-quality estimates of prevalence across traveller subpopulations.

system for targeted COVID-19 screening with the dual goals of identifying asymptomatic, infected travellers and providing real-time information to policymakers for downstream decisions.

Overview of the Eva system

Eva as presented here was deployed across all 40 points of entry to Greece, including airports, land crossings and seaports from 6 August to 1 November 2020. Figure 1 schematically illustrates its operation; Supplementary Fig. 7 provides a more detailed schematic diagram of Eva's architecture and data flow.

We next describe the main steps in processing a passenger.

Passenger locator form

All travellers must complete a passenger locator form (PLF; one per household) at least 24 h before arrival, containing (among other data) information on their origin country, demographics, and point and date of entry. Ref. [10] describes the exact fields and how these sensitive data were handled securely.

Estimating prevalence among traveller types

We estimate traveller-specific COVID-19 prevalence using recent testing results from previous travellers through Eva. Prevalence estimation entails two steps. First, we leverage LASSO (least absolute shrinkage and selection operator) regression from high-dimensional statistics[11] to adaptively extract a minimal set of discrete, interpretable traveller types based on their demographic features (country, region, age and gender); these types are updated on a weekly basis using recent testing results. Second, we use an empirical Bayes method to estimate each type's prevalence daily. Empirical Bayes has previously been used in the body of literature on epidemiology to estimate prevalence across many populations[12,13]. In our setting, COVID-19 prevalence is generally low (for example, ~2 in 1,000), and arrival rates differ substantively across countries. Combined, these features cause our testing data to be both imbalanced (few positive cases among those tested) and sparse (few arrivals from certain countries). Our empirical Bayes method seamlessly handles both challenges. Estimation details are provided in Supplementary Methods 2.2.

Allocating scarce tests

Leveraging these prevalence estimates, Eva targets a subset of travellers for (group) PCR testing on arrival on the basis of their type alone, but no other personal information. The Greek National COVID-19 Committee of Experts approved group (Dorfman) testing[14] in groups of five but eschewed larger groups and rapid testing owing to concerns over testing accuracy.

Eva's targeting must respect various port-level budget and resource constraints that reflect Greece's testing supply chain, which included 400 health workers staffing 40 points of entry, 32 laboratories across the country and delivery logistics for biological samples. These constraints were (exogenously) defined and adjusted throughout the summer by the General Secretariat of Public Health.

The testing allocation decision is entirely algorithmic and balances two objectives. First, given current information, Eva seeks to maximize the number of infected asymptomatic travellers identified (exploitation). Second, Eva strategically allocates some tests to traveller types for which it does not currently have precise estimates to better learn their prevalence (exploration). This is a crucial feedback step. Today's allocations will determine the available data in the prevalence estimation step above when determining future prevalence estimates. Hence, if Eva simply (greedily) sought to allocate tests to types that currently had high prevalence, then, in a few days, it would not have any recent testing data about many other types that had moderate prevalence. Since COVID-19 prevalence can spike quickly and unexpectedly, this would leave a 'blind spot' for the algorithm and pose a serious public health risk. Such allocation problems can be formulated as multi-armed bandits[15–18]—which are widely studied within the body of literature on reinforcement learning—and have been used in numerous applications such as mobile health[19], clinical trial design[20], online advertising[21] and recommender systems[22].

Our application is a nonstationary[23,24], contextual[25], batched bandit problem with delayed feedback[26,27] and constraints[28]. Although these features have been studied in isolation, their combination and practical implementation poses unique challenges. One such challenge is accounting for information from 'pipeline' tests (allocated tests whose results have not yet been returned); we introduce a novel algorithmic technique of certainty-equivalent updates to model information we expect to receive from these tests, allowing us to effectively balance exploration and exploitation in nonstationary, batched settings. To improve interpretability, we build on the optimistic Gittins index for multi-armed bandits[29]; each type is associated with a deterministic index that represents its current 'risk score', incorporating both its estimated prevalence and uncertainty. Algorithm details are provided in Supplementary Methods 2.3.

Grey-listing recommendations

Eva's prevalence estimates are also used to recommend particularly risky countries to be grey-listed, in conjunction with the Greek COVID-19
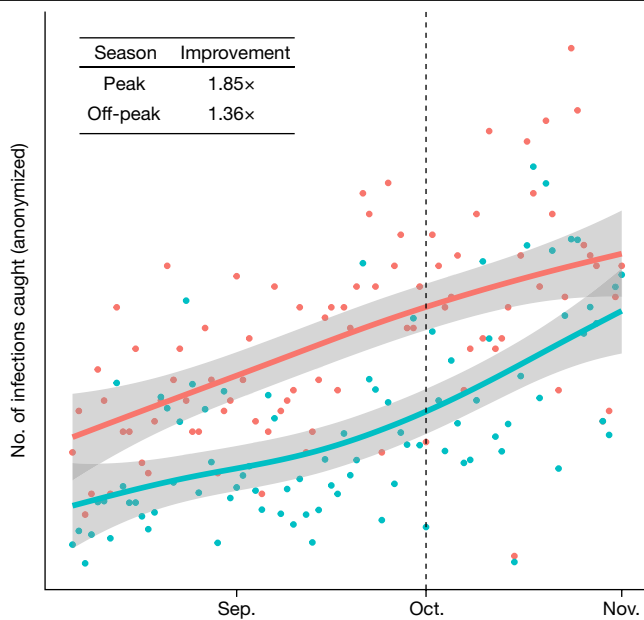
Fig. 2 | **Comparing Eva versus randomized surveillance testing.** The number of infections caught by Eva (red) versus the estimated number of cases caught by random surveillance testing (teal). The peak (respectively, off-peak) season is 6 August to 1 October (respectively, 1 October to 1 November) and is denoted with triangular (respectively, circular) markers. Seasons are separated by the dashed line. The solid lines denote cubic-spline smoothing, with the 95% confidence intervals in grey.

Fig. 3 | **Relative efficacy of Eva over random surveillance versus fraction tested.** The ratio of the number of infections caught by Eva relative to the number of (estimated) infections caught by random surveillance testing, as a function of the fraction of tested travellers. The short-dashed (respectively, long-dashed) line indicates the average fraction tested during the peak (respectively, off-peak) tourist season. Triangular (circular) markers denote estimates from peak (off-peak) days. The solid blue line denotes cubic-spline smoothing, with the 95% confidence interval in grey.

taskforce and the Presidency of the Government. Grey-listing a country entails a tradeoff: requiring a PCR test reduces the prevalence among incoming travellers, but it also reduces non-essential travel significantly (approximately 39%; Supplementary Methods 5), because of the relative difficulty/expense in obtaining PCR tests in summer 2020. Hence, Eva recommends grey-listing a country only when necessary to keep the daily flow of (uncaught) infected travellers at a sufficiently low level to avoid overwhelming contact-tracing teams[30]. Ten countries were grey-listed over the summer of 2020 (Supplementary Methods 5).

Unlike testing decisions, our grey-listing decisions were not fully algorithmic, but instead involved human input. Indeed, while in theory, one might determine an 'optimal' cutoff for grey-listing to balance infected arrivals and reduced travel, in practice it is difficult to elicit such preferences from decision-makers directly. Rather, they preferred to retain some flexibility in grey-listing to consider other factors in their decisions.
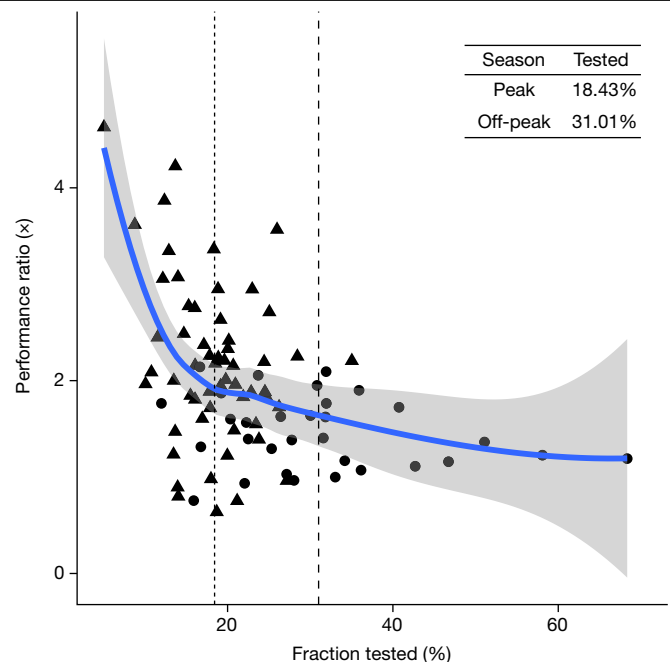
Closing the loop

Results from the tests performed according to the test allocation step are logged within 24–48 h, and then used to update the prevalence estimates from the previous step.

To give a sense of scale, during peak season (August and September), Eva processed 41,830 (±12,784) PLFs each day, and 16.7% (±4.8%) of arriving households were tested each day.

Value of targeted testing

We first present the number of asymptomatic, infected travellers caught by Eva relative to random surveillance testing (that is, where every arrival at a port of entry is equally likely to be tested). Random surveillance testing was Greece's initial proposal and is very common, partly because it requires no information infrastructure to implement. However, we find that such an approach comes at a significant cost to performance (and therefore public health).

We perform counterfactual analysis using inverse propensity weighting[31,32], which provides a model-agnostic, unbiased estimate of the performance of random testing.

During the peak tourist season, we estimate that random surveillance testing would have identified 54.1% (±8.7%) of the infected travellers that Eva identified. (For anonymity, averages and standard deviations are scaled by a (fixed) constant, which we have taken without loss of generality to be the actual number of infections identified by Eva in the same period for ease of comparison.)

In other words, to achieve the same effectiveness as Eva, random testing would have required 85% more tests at each point of entry, a substantive supply chain investment. In October, when arrival rates dropped, the relative performance of random testing improved to 73.4% (±11.0%; Fig. 2). This difference is largely explained by the changing relative scarcity of testing resources (Fig. 3). As arrivals dropped, the fraction of arrivals tested increased, thereby reducing the value of targeted testing. In other words, Eva's targeting is most effective when tests are scarce. In the extreme case of testing 100% of arrivals, targeted testing offers no value as both random and targeted testing policies test everyone. See Supplementary Methods 3 for details.

Value of reinforcement learning

We now compare the performance of Eva with that of policies that require similar infrastructure as Eva, namely PLF data, but instead target testing based on population-level epidemiological metrics (for example, as proposed by the European Union[2]) rather than reinforcement learning. The financial investments of such approaches are similar to those of Eva, and we show that these policies identify fewer cases. (Supplementary Methods 3.2.3 highlights additional drawbacks of these policies, including poor data reliability and a mismatch in prevalence between the general population and the asymptomatic traveller population.)
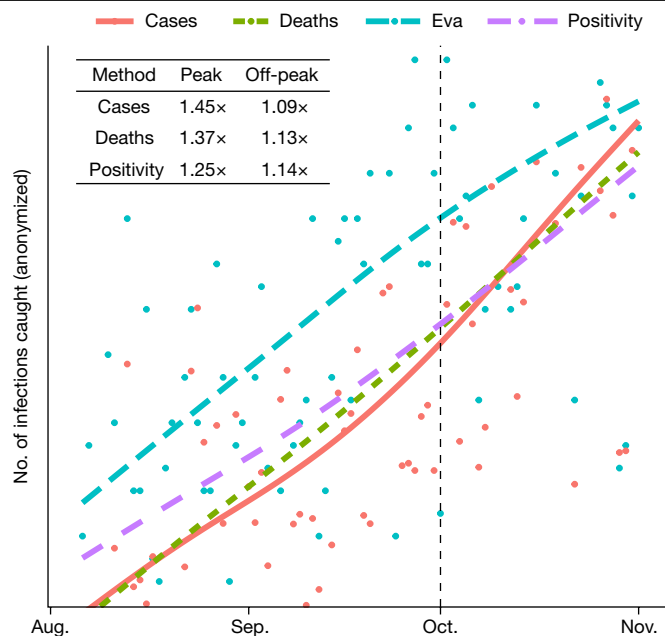
Fig. 4 | Comparing Eva to policies based on epidemiological metrics. The lines represent cubic-spline smoothing of daily infections caught for each policy; raw points are shown only for Eva and the 'Cases' policy for clarity. The dashed line separates the peak (6 August to 1 October) and off-peak (1 October to 1 November) tourist seasons. The inset table shows the relative improvement of Eva over a policy based on the indicated epidemiological metric with the same testing budget for both the peak season and the off-peak season.

Fig. 5 | Predictive power of publicly reported epidemiological metrics. Each of the models 1–4 uses a different subset of features from: 14-day time series of cases per capita, deaths per capita, tests performed per capita and testing positivity rate. Model 5 additionally includes country fixed effects to model country-level idiosyncratic behaviour. Models 1–4 are essentially no better than random prediction, while model 5 achieves slightly better performance. See Supplementary Methods 4.1 for details on model construction and features used in each model. AUROC, area under the receiver operating characteristic curve.
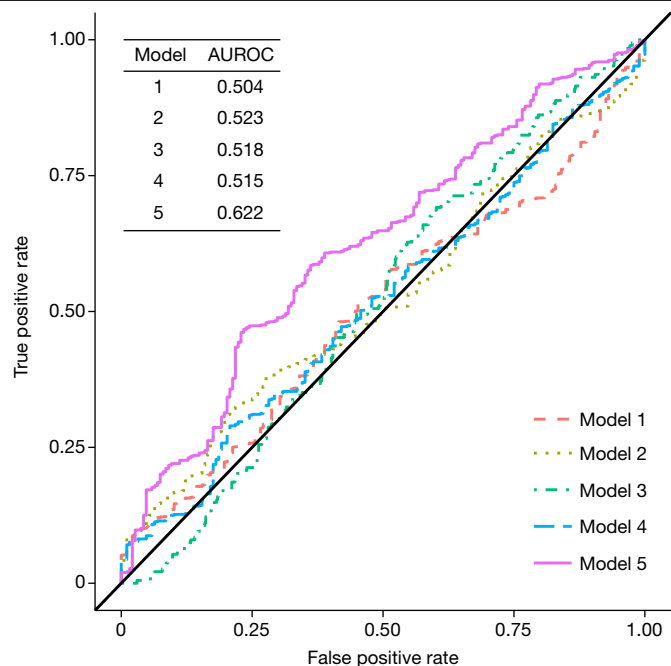
We consider three separate policies that test passengers with probability proportional to cases per capita, deaths per capita or positivity rates for the passenger's country of origin[4–6], while respecting port budgets and arrival constraints. We again use inverse propensity weighting to estimate counterfactual performance (Fig. 4).

During the peak tourist season (August and September), we found that policies based on cases, deaths and positivity rates identified 69.0% (±9.4%), 72.7% (±10.6%) and 79.7% (±9.3%), respectively, of the infected travellers that Eva identified per test. In other words, Eva identified 1.25×–1.45× more infections with the same testing budget and similar PLF infrastructure. In October, when arrival rates dropped, the relative performance of counterfactual policies based on cases, deaths and positivity rates improved to 91.5% (±11.7%), 88.8% (±10.5%) and 87.1% (±10.4%), respectively. Like our results in the previous section, our findings show that the value of smart targeting is larger when testing resources are scarcer. In fact, Eva's relative improvement over these policies was highest in the second half of the peak season (when infection rates were much higher and testing resources were scarcer). See Supplementary Methods 3 for details.

Supplementary Methods 4 discusses possible reasons underlying the poor performance of simple policies based on population-level epidemiological metrics, including reporting delays and systematic differences between the general and asymptomatic traveller populations.

Poor predictive power of epidemiological metrics

Given the poor performance of simple policies based on population-level epidemiological metrics, a natural question is whether more sophisticated functions of these metrics would perform better. Although it is difficult to eliminate this possibility, we argue that this is probably not the case through a related analysis of the extent to which population-level epidemiological metrics can be used to predict COVID-19 prevalence among asymptomatic travellers as measured by Eva. Surprisingly, our findings suggest that widely used epidemiological data are generally ineffective in predicting the actual prevalence of COVID-19 among asymptomatic travellers (the group of interest for border control policies).

Specifically, we examine the extent to which these data can be used to classify a country as high risk (more than 0.5% prevalence) or low risk (less than 0.5% prevalence); such a classification informs whether a country should be grey- or black-listed. (A cutoff of 0.5% was typical for initiating grey-listing discussions with the Greek COVID-19 taskforce, but our results are qualitatively similar across a range of cutoffs.) We compute the true label for a country at each point in time on the basis of Eva's (real-time) estimates. We then train several models using a gradient boosted machine[33] on different subsets of covariates derived from the 14-day time series of cases per capita, deaths per capita, testing rates per capita and testing positivity rates. Figure 5 summarizes their predictive accuracy; we obtained similar results for other state-of-the-art machine learning algorithms.

Note that a random model that uses no data has an area under the receiver operating characteristic curve of 0.5. Thus, models 1–4 offer essentially no predictive value, suggesting that these population-level epidemiological data are not informative of prevalence among asymptomatic travellers.

Model 5, which additionally uses country-level fixed effects, offers some improvement. 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, epidemiological data. The improvement of model 5 suggests that these unobserved drivers are critical to distinguishing high- and low-risk countries.

Overall, this analysis raises concerns not only about travel protocols proposed by the European Union[2] based solely on widely used epidemiological metrics, but also about any protocol that treats all countries

Article

symmetrically. Indeed, the idiosyncratic effects of model 5 suggest that the threshold for deciding whether COVID-19 prevalence in travellers from country A is spiking may differ significantly from that of country B. See Supplementary Methods 4.1 for details.

In Supplementary Methods 4.3, we also study the information delay between a country's publicly reported cases (the most common metric) and prevalence among asymptomatic travellers from that country. We expect a lag because of the time taken for symptoms to manifest, and reporting delays induced by poor infrastructure. We find a modal delay of 9 days.

Value of grey-listing

Eva's measurements of COVID-19 prevalence were also used to provide early warnings for high-risk regions, in response to which Greece adjusted travel protocols by grey-listing these nations. We estimate that Eva prevented an additional 6.7% (±1.2%) infected travellers from entering the country through its early grey-listing decisions in the peak season; results in the off-peak season are similar. For privacy, we have expressed the benefit relative to the number of infected travellers identified by Eva. See Supplementary Methods 5 for details.

Lessons learned from deployment and design

Eva is a large-scale data-driven system that was designed and deployed during the COVID-19 crisis. Leading up to and throughout deployment, we met twice a week with the COVID-19 Executive Committee of Greece, an interdisciplinary team of scientists and policymakers. Through those meetings, we gleaned several lessons that shaped Eva's design and contributed to its success.

Design the algorithm around data minimization

Data minimization (that is, requesting the minimum required information for a task) is a fundamental tenet of data privacy and the General Data Protection Regulation (GDPR). We met with lawyers, epidemiologists and policymakers before designing the algorithm to determine what data and granularity may legally and ethically be solicited by the PLF. Data minimization naturally entails a tradeoff between privacy and effectiveness. We limited requests to features thought to be predictive on the basis of the best available research at the time (origin, age and gender[34,35]), but omitted potentially informative but invasive features (for example, occupation). We further designed our empirical Bayes estimation strategy around these data limitations.

Prioritize interpretability

For all parties to evaluate and trust the recommendations of a system, the system must provide transparent reasoning. An example from our deployment was the need to communicate the rationale for 'exploration' tests (that is tests for types with moderate but very uncertain prevalence estimates). Such tests may seem wasteful. Our choice of empirical Bayes allowed us to easily communicate that types with large confidence intervals may have significantly higher risk than their point estimate suggests, and thus require some tests to resolve uncertainty (see, for example, Supplementary Figs. 9 and 11, which were featured on policymakers' dashboards).

A second example was our choice to use Gittins indices, which provide a simple, deterministic risk metric for each type that incorporates both estimated prevalence and corresponding uncertainty, driving intuitive test allocations. In contrast, using upper-confidence-bound or Thompson sampling with logistic regression[36,37] would have made it more difficult to visualize uncertainty (a high-dimensional ellipsoid or posterior distribution), and test allocations would depend on this uncertainty through an opaque computation (a high-dimensional projection or stochastic sampling).

This transparency fostered trust across ministries of the Greek Government using our estimates to inform downstream policymaking, including targeting contact-tracing teams, staffing of mobile testing units and adjusting social distancing measures.

Design for flexibility

Finally, as these systems require substantial financial and technical investment, they need to be flexible to accommodate unexpected changes. We designed Eva in a modular manner disassociating type extraction, estimation and test allocation. Consequently, one module can easily be updated without altering the remaining modules. For example, had vaccine distribution begun in the summer of 2020, we could define new types based on passengers' vaccination status without altering our procedure for prevalence estimates or test allocation. Similarly, if rapid testing were approved, our allocation mechanism could be updated to neglect delayed feedback without affecting other components. This flexibility promotes longevity, as it is easier to get stakeholder buy-in for small adjustments to an existing system than for a substantively new approach.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-021-04014-z.

1. Council of the European Union. *Council Recommendation on the Temporary Restriction on Non-essential Travel into the EU and the Possible Lifting of Such Restriction* https://www.consilium.europa.eu/media/47592/st_9208_2020_init_en.pdf (30 June 2020).
2. European Commission. *Draft Council Recommendation on a Coordinated Approach to the Restriction of Free Movement in Response to the COVID-19 Pandemic* (12 October 2020).
3. World Travel and Tourism Council. *Recovery Scenarios 2020 & Economic Impact from COVID-19* https://wttc.org/Research/Economic-Impact/Recovery-Scenarios (November 2020).
4. Hasell, J. et al. A cross-country database of COVID-19 testing. *Sci. Data* **7**, 345 (2020).
5. Roser, M., Ritchie, H., Ortiz-Ospina, E. & Hasell, J. *Coronavirus Pandemic (COVID-19)* https://ourworldindata.org/coronavirus (OurWorldInData.org, 2020).
6. Dong, E., Du, H. & Gardner, L. An interactive web-based dashboard to track COVID-19 in real time. *Lancet Infect. Dis.* **20**, 533–534 (2020).
7. Wu, S. L. et al. Substantial underestimation of SARS-CoV-2 infection in the United States. *Nat. Commun.* **11**, 4507 (2020).
8. Cevik, M. et al. SARS-CoV-2, SARS-CoV, and MERS-CoV viral load dynamics, duration of viral shedding, and infectiousness: a systematic review and meta-analysis. *Lancet Microbe* **2**, e13–e22 (2021).
9. Phipps, S., Grafton, Q. & Kompas, T. Robust estimates of the true (population) infection rate for COVID-19: a backcasting approach. *R. Soc. Open Sci.* **7**, 200909 (2020).
10. Ministry of Civil Protection and Ministry of Tourism, Hellenic Republic. *Protocol for Arrivals in Greece* https://travel.gov.gr/#/policy (1 July 2020).
11. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. B* **58.1**, 267–288 (1996).
12. Greenland, S. & Robins, J. Empirical-Bayes adjustments for multiple comparisons are sometimes useful. *Epidemiology* **2**, 244–251 (1991).
13. Devine, O. J., Louis, T. & Halloran, E. Empirical Bayes methods for stabilizing incidence rates before mapping. *Epidemiology* **5**, 622–630 (1994).
14. Dorfman, R. The detection of defective members of large populations. *Ann. Math. Stat.* **14**, 436–440 (1943).
15. Thompson, W. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika* **25.3/4**, 285–294 (1933).
16. Lai, T. L. & Robbins, H. Asymptotically efficient adaptive allocation rules. *Adv. Appl. Math.* **6.1**, 4–22 (1985).
17. Gittins, J. Bandit processes and dynamic allocation indices. *J. R. Stat. Soc. B* **41.2**, 148–164 (1979).
18. Auer, P. Using confidence bounds for exploitation-exploration trade-offs. *J. Mach. Learn. Res.* **3**, 397–422 (2002).
19. Tewari, A. & Murphy, S. A. In *Mobile Health—Sensors, Analytic Methods, and Applications* (eds Rehg, J. M., Murphy, S. A. & Kumar, S.) 495–517 (Springer, 2017).
20. Durand, A. et al. Contextual bandits for adapting treatment in a mouse model of de novo carcinogenesis. In *Proceedings of the 3rd Machine Learning for Healthcare Conference. Proc. Mach. Learn. Res.* **85**, 67–82 (2018).
21. Li, L., Chu, W., Langford, J. & Schapire, R. In *Proceedings of the 19th International Conference on World Wide Web* 6611–6670 (Association for Computing Machinery, 2010); https://dl.acm.org/doi/10.1145/1772690.1772758
22. Amat, F., Chandrashekar, A., Jebara, T. & Basilico, J. In *Proceedings of the 12th ACM Conference on Recommender Systems* 487–488 (Association for Computing Machinery, 2018); https://dl.acm.org/doi/10.1145/3240323.3241729

23. Besbes, O., Gur Y. & Zeevi, A. In *Advances in Neural Information Processing Systems 27* (eds Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. & Weinberger, K. Q.) 199–207 (NIPS, 2014).
24. Luo, H., Wei, C.-Y., Agarwal, A. & Langford, J. Efficient contextual bandits in non-stationary worlds. In *Proceedings of the 31st Conference on Learning Theory. Proc. Mach. Learn. Res.* **75**, 1739–1776 (2018).
25. Bastani, H. & Bayati, M. Online decision making with high-dimensional covariates. *Oper. Res.* **68.1**, 276–294 (2020).
26. Gao, Z., Han, Y., Ren, Z. & Zhou, Z. In *Advances in Neural Information Processing Systems 32* (eds Wallach, H. et al.) 503–514 (NeurIPS, 2019).
27. Perchet, V., Rigollet, P., Chassang, S. & Snowberg, E. Batched bandit problems. *Ann. Stat.* **44.2**, 660–681 (2016).
28. Agrawal, S. & Devanur, N. Bandits with concave rewards and convex knapsacks. In *Proceedings of the 15th ACM Conference on Economics and Computation* 989–1006 (Association for Computing Machinery, 2014); https://doi.org/10.1145/2600057.2602844
29. Gutin, E. & Farias, V. In *Advances in Neural Information Processing Systems 29* (eds Lee, D., Sugiyama, M., Luxburg, U., Guyon, I. & Garnett, R.) 3153–3161 (2016).
30. Hellewell, J. et al. Feasibility of controlling COVID-19 outbreaks by isolation of cases and contacts. *Lancet Glob. Health* **8**, e488–e496 (2020).
31. Imbens, W. G. & Rubin, B. D. *Causal Inference in Statistics, Social and Biomedical Sciences* (Cambridge Univ. Press, 2015).
32. Rosenbaum, P. & Rubin, D. The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41–55 (1983).
33. Friedman, J. H. Greedy function approximation: a gradient boosting machine. *Ann. Stat.* **29**, 1189–1232 (2001).
34. Davies, N. et al. Age-dependent effects in the transmission and control of COVID-19 epidemics. *Nat. Med.* **26**, 1205–1211 (2020).
35. Davies, S. E. & Bennet, B. A gendered human rights analysis of Ebola and Zika: locating gender in global health emergencies. *Int. Aff.* **92.5**, 1041–1060 (2016).
36. Agrawal S. & Goyal, N. Thompson sampling for contextual bandits with linear payoffs. In *Proceedings of the 30th International Conference on Machine Learning. Proc. Mach. Learn. Res.* **28**, 127–135 (2013).
37. Chu, W., Li, L., Reyzin, L. & Schapire, R. Contextual bandits with linear payoff functions. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics. Proc. Mach. Learn. Res.* **15**, 208–214 (2011).

Article

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

To support further research, aggregated, anonymized data are available at https://github.com/kimondr/EVA_Public_Data. These data aggregate passenger arrival and testing information over pairs of consecutive days, country of origin and point of entry. The finer granularity data that support the (exact) findings of this study are protected by GDPR. These data are available from the Greek Ministry of Civil Protection, but restrictions apply to the availability of these data, which were used under licence for the current study, and so are not publicly available. Access to these data can be granted by the Greek Ministry of Civil Protection (info@gscp.gr) only for research that is conducted in the public interest for public health (GDPR Recital 159) and scientific purposes (GDPR Article 89). Finally, the population-level epidemiological metrics used in our analysis can be obtained freely from the Our World In Data COVID-19 dataset (https://github.com/owid/covid-19-data/tree/master/public/data).

Code availability

All code used in this paper was written in a combination of R and Python 3.7. The code for the deployment of the algorithm on a sample dataset is available at https://github.com/vgupta1/EvaTargetedCovid19Testing. The code for reproducing the results of our counterfactual analysis is available at https://github.com/vgupta1/Eva_CounterfactualAnalysis.

Additional information

nature portfolio

Corresponding author(s):   Kimon Drakopoulos

Last updated by author(s): Sep 3, 2021

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Our system was deployed on Amazon Web Services (AWS). Code written in Python 3.7 and R version 4.0.5. The code for the algorithm and for the Counterfactual Analysis is posted on Github: https://github.com/vgupta1/EvaTargetedCovid19Testing and at https://github.com/vgupta1/Eva_CounterfactualAnalysis, respectively.
Data analysis	Data Analysis was performed using Python 3.7 and R version 4.0.5

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:
- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

To support further research, aggregated, anonymized data are available at https://github.com/kimondr/EVA_Public_Data. These data aggregate passenger arrival and testing information over pairs of consecutive days, country of origin, and point of entry.

The finer granularity data that support the (exact) findings of this study are protected by GDPR. These data are available from the Greek Ministry of Civil Protection but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Access to these data

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences   ☐ Behavioural & social sciences   ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	On average, 38,500 (±13,590) PLFs were processed each day (with higher traffic in August and September), and 18.3% (± 6.1%) of arriving households per day were tested (accounting for no-shows). The effective sample size in our context was determined by the testing availability at each point of entry and was pre-specified by the General Secretariat of Public Health at the Greek Ministry of Health. As evidenced by our confidence intervals, these sample sizes were indeed sufficient.
Data exclusions	0.6% of arrivals were tested by border agents at their own discretion (i.e., not recommended by Eva) and we drop these 0.6% of arrivals from the analysis. The exclusion criteria were pre-established.
Replication	We found that our algorithm consistently improved performance compared to benchmark policies throughout the summer. However, replication beyond the Greek border control context is out of the scope of our study.
Randomization	Samples were allocated based on our reinforcement learning algorithm. A detailed description is provided in Methods.
Blinding	Blinding is not relevant to this study. Border agents performed tests on individuals and (pseudonymized) samples were sent to laboratories that produced the result (positive or negative) for each traveler. The investigators (authors of this paper) were only provided pseudonymized, aggregate results.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging