Deploying an Artificial Intelligence System for COVID-19 Testing at the Greek Border

Hamsa Bastani*[,1], Kimon Drakopoulos*[,2,†], Vishal Gupta*[,2], Jon Vlachogiannis[3],

Christos Hadjicristodoulou[4], Pagona Lagiou[5], Gkikas Magiorkinis[5], Dimitrios Paraskevis[5], Sotirios Tsiodras[6]

[1]Department of Operations, Information and Decisions, Wharton School, University of Pennsylvania
[2]Department of Data Sciences and Operations, Marshall School of Business, University of Southern California
[3]AgentRisk
[4]Department of Hygiene and Epidemiology, University of Thessaly
[5]Department of Hygiene, Epidemiology and Medical Statistics, School of Medicine, National and Kapodistrian University of Athens
[6] Department of Internal Medicine, Attikon University Hospital, Medical School, National and Kapodistrian University of Athens, Greece
*These authors contributed equally to this work
† Corresponding author. Email: drakopou@marshall.usc.edu

On July 1st, 2020, members of the European Union gradually lifted earlier COVID-19 restrictions on non-essential travel. In response, we designed and deployed "Eva" – a novel, self-learning artificial intelligence system – across all Greek borders to identify asymptomatic travelers infected with SARS-CoV-2 based on demographic characteristics and results from previously tested travelers. Eva allocates Greece's limited testing resources to (i) limit the importation of new cases and (ii) provide real-time estimates of COVID-19 prevalence to inform border policies.

Counterfactual analysis shows that our system identified on average 1.85x as many asymptomatic, infected travelers as random surveillance testing, and up to 2-4x as many during peak travel. Moreover, for most countries, Eva identified atypically high prevalence 9-days earlier than machine learning systems based on publicly reported data. By adaptively adjusting border policies 9-days earlier, Eva prevented additional infected travelers from arriving.

Finally, using Eva's unique cross-country, large-scale dataset on prevalence in asymptomatic populations, we show that commonly used public data on cases/deaths/testing have limited predictive value for the actual prevalence among asymptomatic travelers, and furthermore exhibit strong country-specific idiosyncrasies. As herd immunity is still likely more than a year away [1]**, and travel protocols for the summer of 2021 are still being discussed, our insights raise serious concerns about internationally proposed border control policies** [2] **that are both country-agnostic and solely based on public data. Instead, our work paves the way for leveraging AI and real-time data for public health goals, such as border control during a pandemic.**

Introduction

In the first wave of the pandemic, many countries restricted non-essential travel to mitigate the spread of COVID-19. The restrictions crippled most tourist economies, with estimated losses of 1 trillion USD 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 labor. Yet, safeguarding public health naturally requires limiting the number of SARS-CoV-2 infected travelers entering a country. Ideally, border agents would test every arriving traveler and require those that test positive and their contacts to quarantine [4, 5, 6]; however, the scarcity of testing resources over the summer of 2020 made this approach (even with group testing [7]) untenable. As a result, targeting safety measures towards "high-risk" traveler profiles became necessary [8].

Different countries adopted different border screening protocols, typically based upon the origin country of the traveler. Despite their variety, we group the protocols used in early summer 2020 into 4 broad types:

- Allowing unrestricted travel from designated "white-list" countries.
- Requiring travelers from designated "grey-listed" countries to provide proof of a negative RT-PCR test before arrival.
- Requiring all travelers from designated "red-listed" countries to quarantine upon 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 "color" to assign to a traveler differed across nations. For example, as of July 1st, 2020, Spain designated the countries specified in [9] as white-listed, and black-listed all other countries while, in contrast, Croatia designated all countries in [9] as grey-listed or red-listed (traveler's choice).

To the best of our knowledge, in all European nations except Greece, the above "color designations" were entirely based on *publicly available data* (e.g., see [9, 2]), and, in particular, country-level data on 14-day rate of cases, deaths, and/or testing that were available in the public domain [10, 11, 12].[1]

However, publicly reported COVID-19 data are imperfect. Different countries follow different reporting protocols and testing strategies. Significant underreporting is known to occur, and the extent differs significantly across regions [13]. Moreover, many strained testing systems focus testing on symptomatic and severely ill patients. Reported prevalence in these *symptomatic* groups may not reflect the prevalence among pre-symptomatic or asymptomatic travelers, particularly because epidemiological research suggests pre-symptomatic or asymptomatic individuals may shed the virus several days after infection [14, 15, 16]. Furthermore, public data generally suffers reporting delays, e.g., due to poor infrastructure.

These drawbacks of publicly reported epidemiological data motivated our design and nation-wide deployment of Eva: the first fully algorithmic, real-time and self-learning AI system for targeted COVID-19 screening.

The Eva System: Overview

Fig. 1 schematically illustrates the operation and flow of information for Eva.

[1] An exception is the UK, which engaged in small-scale testing at select airports that *may* have informed their policies.

1. Passenger Locator Form (PLF)
All travelers must complete a PLF (one per household) at least 24 hours prior to arrival, containing information on their origin country, demographics, point and date of entry, and intended destination.

2. Estimating Prevalence among Traveler Types
Using recent testing results from previous travelers through Eva, we estimate the COVID-19 prevalence among different types of passengers. This entails two steps: First, we leverage techniques from high-dimensional statistics [17] to identify types of passengers (e.g., travelers from Madrid, Spain) that likely share similar prevalence. Second, we use an empirical Bayes method to estimate each type's prevalence. Empirical Bayes has been previously used in the epidemiological literature to estimate prevalence across many populations [18, 19]. In our setting, COVID-19 prevalence is generally low (e.g., ~2 in 1000), 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 Section 2.2 of Methods.

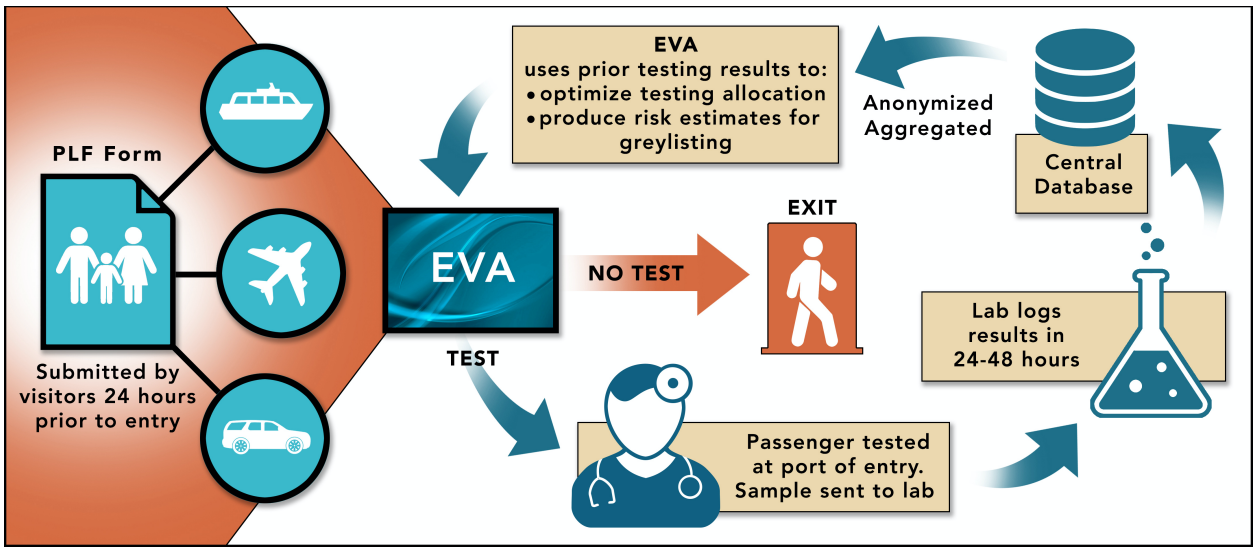


Figure 1: Project Eva – a self-learning artificial-intelligence system for COVID-19 screening that we deployed during the summer of 2020 across all 40 points of entry to Greece, including airports, land crossings, and seaports.

3. Allocating Scarce Tests
Leveraging these prevalence estimates, Eva targets a subset of travelers for (group) PCR testing upon arrival based on their (anonymized) PLF data.[2] This 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 targeting decision is entirely algorithmic and balances two objectives: First, given current information, Eva seeks to maximize the number of asymptomatic, infected travelers identified. Second, Eva strategically allocates some tests to traveler types for which it does not currently have precise

[2] National COVID-19 Committee of Experts approved group (Dorfman) testing [7] with a maximum group size of 5. Larger groups and rapid testing were not approved at the time due to concerns over testing accuracy.

prevalence estimates. *This is a crucial feedback step.* Today's allocations will determine the available data in Step 2 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.

Said differently, Eva must balance the well-studied tradeoff between *exploration* (learning prevalence for many types) and *exploitation* (finding infected travelers) [8]. Such allocation problems are widely studied in operations research, computer science and statistics in the multiarmed bandit literature [20, 21, 22, 23, 24] and have been used in numerous applications such as mobile health [25], clinical trial design [26], A/B testing [27], online advertising [28], recommender systems [29], customer support [30], and targeted assistance for refugees [31]. Our particular application is a nonstationary [32, 33], contextual [34], batched bandit problem with delayed feedback [35, 36] and constraints [37]. Although these features have been studied in isolation, their combination poses unique challenges. We propose a novel algorithm that balances the exploration-exploitation tradeoff in this setting, while respecting budget/arrival constraints at each point of entry. Algorithm details are provided in Section 2.3 of Methods.

4. Grey-listing Recommendations

Eva also uses current prevalence estimates to recommend particularly risky countries to be grey-listed. Clearly, grey-listing *all* countries would minimize incoming infections, but this would also entail a substantive drop in non-essential travel (~39%, as shown in Section 3.2 of Methods), incurring substantial economic costs. Hence, Eva recommends grey-listing a country only when necessary to keep the daily flow of (uncaught) infected travelers at a sufficiently low level, to avoid overwhelming downstream contact-tracing teams and healthcare systems [38]. Grey-listing recommendations were made in conjunction with the Greek COVID-19 taskforce, and further reviewed/authorized by the Presidency of the Government. Ten countries were grey-listed over the summer of 2020.

5. Closing the Loop

Anonymized results from tests performed in Step 3 are processed and added to our database in 24-48 hours, and then used to update our prevalence estimates in Step 2.

Eva as presented above was in operation from August 6th to November 1st.[3] Eva monitored 40 points of entry, including airports, ports and land crossings. On average, 38,500 (±13,590) PLFs were processed *each day*, with higher traffic in August and September. Not all households who submitted a PLF arrived at the border – the daily "no-show" rate was 17.6% (± 6.9%). On average, budget constraints allowed testing 18.3% (± 6.1%) of arriving households per day, with a smaller fraction in high-traffic months.

[3] From July 1st to August 6th, while appropriate infrastructure was built, an "open loop" version of Eva was used, which is not discussed in this paper.

Results

Value of Targeting: Comparison to Random Surveillance Testing

We first present the number of asymptomatic, 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]. Importantly, IPW provides a *model-agnostic, unbiased* estimate of performance – i.e., our estimate of random surveillance testing's performance is unbiased *regardless* of the accuracy of our COVID-19 prevalence estimates. This makes IPW an ideal method for benchmarking Eva's performance.

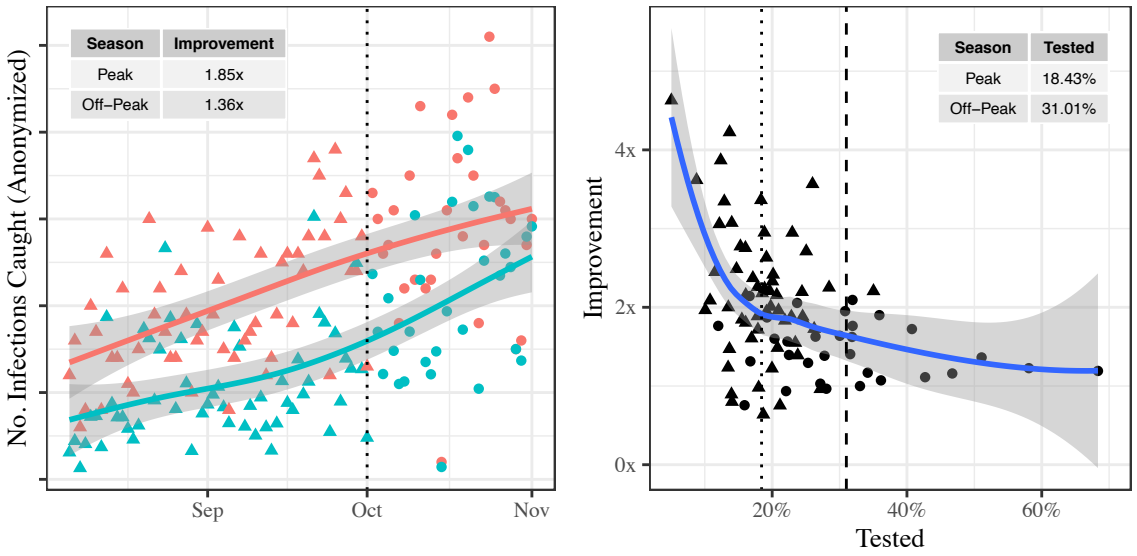


Figure 2 Left: Infections caught by our policy (red) vs estimated number of cases caught by random, surveillance testing (teal). The peak (resp. off-peak) season is Aug. 6 to Oct. 1 (resp. Oct. 1 to Nov. 1) is denoted with triangular (resp. circular) markers. Seasons separated by dotted line. Solid lines denote LOESS smoothing with 95% confidence intervals in grey. Right: Ratio of our policy's daily improvement over random, surveillance testing vs the fraction of tested travelers. Dotted line indicates the average fraction tested during the peak tourist season. Triangular (circular) markers denote estimates from peak (off-peak) days. Solid blue line denotes LOESS smoothing with 95% confidence interval in grey.

During the peak tourist season (August, September), we find that Eva identified 1.85x (±.06) as many asymptomatic, infected travelers as random surveillance testing.[4] In other words, to achieve the same effectiveness as Eva, random testing would have required 85% more tests at each point of entry. In October, when arrival rates dropped, this relative improvement dropped to 1.36x (±.05) (see Fig. 2, left panel). This difference is largely explained by the changing relative scarcity of testing resources (see Fig. 2, right panel). 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. Details are provided in Section 3.1 of Methods.

[4] For anonymity, standard errors and averages are presented relative to the (estimated) average performance of random surveillance testing.

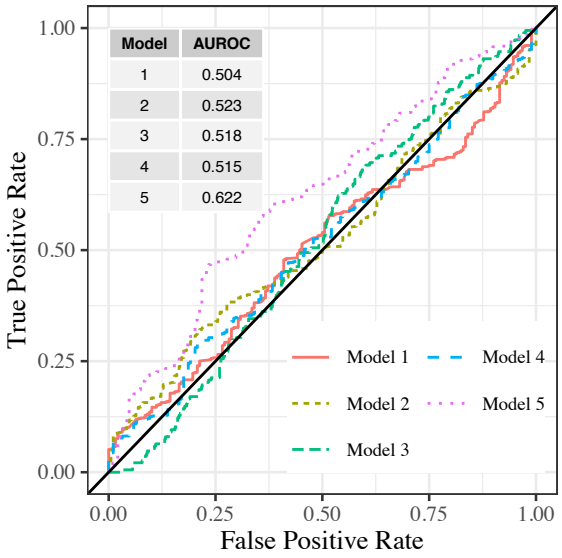
Value of Surveillance: Poor Predictivity and Information Delays in Public Data

Recently, the EU [2] proposed testing protocols for non-essential travel where the color designation of a country is determined from commonly-used, publicly reported epidemiological data: specifically, country-level 14-day rate of cases, deaths, and/or tests administered [10, 11, 12]. This color designation intends to capture the actual prevalence among travelers from each country, so that restrictions to free movement are imposed only when necessary to protect public health.

As discussed earlier, such data are affected by country-specific testing protocols (e.g., testing symptomatic patients versus extensive population-wide testing) and, thus, these data may not accurately reflect the actual prevalence among *asymptomatic* travelers (the group of interest for border control policies).

The data collected via Eva provide the first, large-scale dataset on asymptomatic populations across nations. This allows us to assess the extent to which different countries' publicly reported data can be used to estimate true prevalence. To the best of our knowledge, this is the first study of this kind. Surprisingly, our findings suggest that publicly reported epidemiological data are generally *ineffective* in estimating the actual prevalence of COVID-19 among asymptomatic travelers, and, hence, travel protocols based only on these data (like those proposed by the EU in [2]) may be inadequate.

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); this classification problem is key to determining whether a country should be grey- or black-listed.[5] We compute the true label for a country at each point in time based on Eva's (real-time) estimates. We build several models based upon different covariates, training each with a Gradient Boosted Machine (GBM),[6] as shown in Fig. 3. Details are provided in Section 4.1 of Methods.



Model	Features Used for Prediction
1	14-day average of cases and deaths per capita
2	14-day average of cases, deaths and tests performed per capita, and the positivity rate
3	14-day timeseries of cases and deaths per capita
4	14-day timeseries of cases, deaths and tests performed per capita, and the positivity rate
5	14-day timeseries of cases, deaths and tests performed per capita, the positivity rate, and country fixed effects

[5] 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.

[6] GBM is a machine-learning algorithm based on tree ensembles, which are known to perform well in structured classification tasks [50, 48]. We obtain similar results with other algorithms such as RNNs or LASSO.

Figure 3: Predictive power of publicly reported epidemiological data. "14-day" time series indicates the previous 14 days of the variable were used as features. Models 1-4 offer predictions that are essentially no better than random prediction. Model 5 includes country fixed-effects to model country-level idiosyncratic behavior and achieves slightly better predictive power.

Note that a random model that uses no data has an AUROC of 0.5. Thus, Models 1-4 offer essentially no predictive value, suggesting that public data are not informative of actual asymptomatic prevalence. In particular, although publicly reported data have been shown to effectively forecast other publicly reported metrics like cases/deaths/hospitalizations in the US [41], we find that it is surprisingly *ineffective* at forecasting the relevant metric for border control (prevalence among asymptotic travelers).

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 data. Note that the coefficients of these fixed effects can only be inferred using Eva's surveillance testing results (the response variable), making it impossible to implement Model 5 without Eva's testing strategy. The improvement imbued by Model 5 suggests that these unobserved drivers are critical to distinguishing high- and low-risk countries.

Overall, this analysis not only raises concerns about travel protocols proposed by the EU [2] based upon public data, but also about *any* protocol that treats all countries symmetrically. Indeed, the idiosyncratic effects of Model 5 suggest that the thresholds for deciding whether COVID-19 prevalence in travelers from Country A is spiking might differ significantly from that of Country B.

Next, we study the information delay between a country's publicly reported cases (the most commonly used metric) and prevalence among asymptomatic travelers from that country. We expect case data to lag asymptomatic prevalence for two reasons: first, epidemiologically, symptoms can manifest up to 10-17 days after patients are infected [14, 15], and second, (lack of) information infrastructure induces reporting delays.

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. Intuitively, if a country's case data lags its true prevalence by ℓ days, then we should be able to more effectively predict its risk status y_t on day t using time series case data from ℓ days *in the future,* i.e., data from $[t, t+\ell]$. Using this approach, we identify 3 clusters of countries: those with short delays (1 day), medium delays (9 days) and long delays (16 days). The modal value is 9 days. Details are provided in Section 4.2 of Methods.

Value of Early Warning: Cases Prevented

As evidenced by the results above, one benefit of Eva is near-real time measurements of COVID-19 prevalence. These measurements provided early warnings for high-risk regions, in response to which Greece adjusted travel protocols by grey-listing these nations. We next quantify the benefit of these grey-listing decisions.

Note that grey-listing a country has two effects: First, measured prevalence among travelers drops sharply. Second, due to the cost, unavailability, and inconvenience of PCR testing, the number of arrivals also drops. Fig. 4 illustrates both effects for Malta, which was grey-listed on August 12, 2020.

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 models on pre-intervention data from both grey-listed (treated) and non-grey-listed (untreated) countries (see Section 3.2 of Methods for details). Fig. 4 shows the resulting counterfactual estimates for Malta, a representative grey-listed country. As desired, the models visually track both true prevalence and true arrival rates well prior to the grey-listing intervention (overlapping lines).

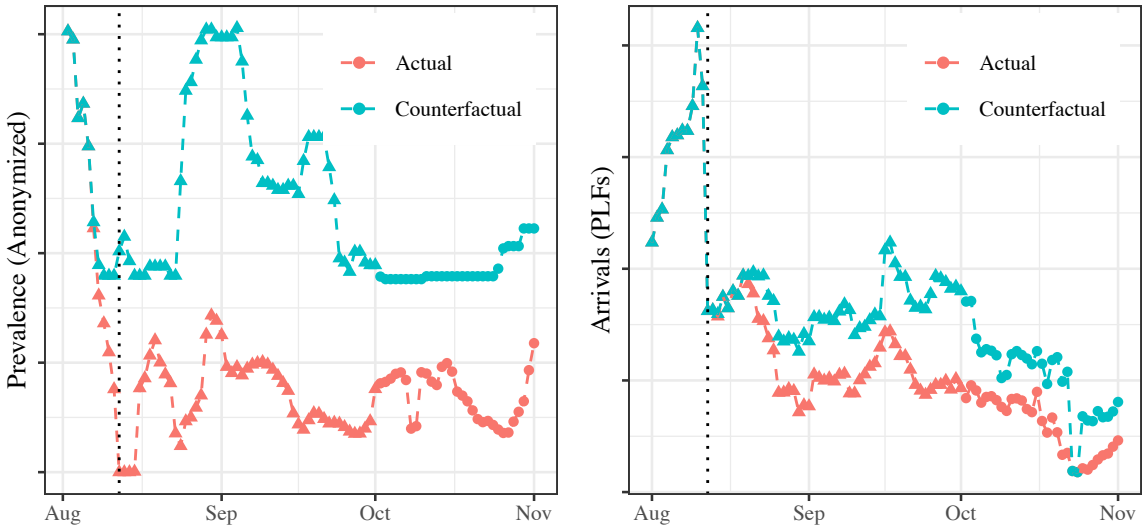


Figure 4: Measured and counterfactually estimated prevalence (left) and arrival rates (right) for Malta before and after-grey-listing. Grey-listing occurred on 12 Aug (dotted line).

To quantify the benefit of early-warnings, we assume that any country that Eva recommended to grey-list would *also* have been grey-listed by a baseline surveillance system, but 9 days later (since our earlier analysis suggests at least a 9-day information delay in public case data for most countries). We then estimate the number of incoming infections from these countries to Greece in these 9 days. We find that during the peak season, Eva prevented an additional .12x (±.022) asymptomatic, infected travelers from entering through its early grey-listing decisions, where we have expressed the benefit (and standard error) relative to the number of infected travelers caught by random surveillance testing. Combined with our earlier estimates, Eva thus prevented roughly 1.97x as many infected travelers from entering as random testing. In the off-peak season, the value of early warnings drops to .09x (±.007), and the total benefit to 1.45x relative to random testing.

Conclusions

To the best of our knowledge, our work is the first to design a novel AI and optimization system for targeted screening and deploy it at a national scale. We illustrate the benefits of such a data-driven strategy for controlling the number of infected travelers entering a country as well as monitoring near real-time prevalence across populations. Through better allocations of limited testing resources and adaptive border control policies, we essentially double the effectiveness of random testing.

Bibliography

[1] WHO, "COVID-19 Virtual Press conference transcript - 11 January 2021," 11 January 2021. [Online]. Available: https://www.who.int/publications/m/item/covid-19-virtual-press-conference-transcript---11-january-2021. [Accessed 1 February 2021].

[2] Draft Council Recommendation on a coordinated approach to the restriction of free movement in response to the COVID-19 pandemic, Brussels, 12 October 2020: General Secretariat of the Council.

[3] "World Travel and Tourism Council," November 2020. [Online]. Available: Council https://wttc.org/Research/Economic-Impact/Recovery-Scenarios.

[4] N. Augenblick, J. T. Kolstad, Z. Obermeyer and A. Wang, "Group testing in a pandemic: The role of frequent testing, correlated risk, and machine learning," *National Bureau of Economic Research,* 2020.

[5] P. Frazier, Y. Zhang and M. Cashore, "Feasibility of COVID-19 Screening for the U.S. Population with Group Testing," Cornell University Working Paper, Ithaca, 2020.

[6] E. Kaplan and H. Forman, "Logistics of aggressive community screening for coronavirus 2019," *JAMA Health Forum,* vol. 1, no. 5, p. e200565, 2020.

[7] R. Dorfman, "The detection of defective members of large populations," *The Annals of Mathematical Statistics,* vol. 14, no. 4, pp. 436-440, 1943.

[8] M. Kasy and A. Teytelboym, "Adaptive targeted infectious disease testing," *Oxford Review of Economic Policy,* vol. 36, pp. S77-S93, 2020.

[9] Council recommendation on the temporary restriction on non-essential travel into the EU and the possible lifting of such restriction, Brussels, 30 June 2020: https://www.consilium.europa.eu/media/47592/st_9208_2020_init_en.pdf.

[10] J. Hasell, E. Mathieu, D. Beltekian, B. a. G. C. a. O.-O. E. Macdonald, M. Roser and H. Ritchie, "A cross-country database of COVID-19 testing," *Scientific data,* vol. 7, no. 1, pp. 1-7, 2020.

[11] M. Roser, H. Ritchie, E. Ortiz-Ospina and J. Hasell, "Coronavirus Pandemic (COVID-19)," OurWorldInData.org, 2020. [Online]. Available: https://ourworldindata.org/coronavirus.

[12] E. Dong, H. Du and L. Gardner, "An interactive web-based dashboard to track COVID-19 in real time," *The Lancet infectious diseases,* vol. 20, no. 5, pp. 533-534, 2020.

[13] S. L. Wu, A. N. Mertens, Y. S. Crider, A. Nguyen, N. N. Pokpongkiat, S. Djajadi, A. Seth, M. S. Hsiang, J. M. J. Colford, A. Reingold, B. F. Arnold, A. Hubbard and J. Benjamin-Chung, "Substantial underestimation of SARS-CoV-2 infection in the United States," *Nature Communications,* 2020.

[14] P. Zhao, L. Zhang, Y. Jiang and Z.-H. Zhou, "A simple approach for non-stationary linear bandits," *International Conference on Artificial Intelligence and Statistics,* pp. 746-755, 2020.

[15] C. Muge, T. Matthew, L. Ollie, A. E. Maraolo, J. Schafers and H. Antonia, "SARS-CoV-2, SARS-CoV, and MERS-CoV viral load dynamics, duration of viral shedding, and infectiousness: a systematic review and meta-analysis," *The Lancet Microbe,* 2020.

[16] S. Phipps, Q. Grafton and T. Kompas, "Robust estimates of the true (population) infection rate for COVID-19: a backcasting approach," *Royal Society Open Science,* vol. 7, no. 11, 2020.

[17] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological),* pp. 267-288, 1996.

[18] S. Greenland and J. Robins, "Empirical-Bayes adjustments for multiple comparisons are sometimes useful," *Epidemiology,* pp. 244-251, 1991.

[19] O. J. Devine, T. Louis and E. Halloran, "Empirical Bayes methods for stabilizing incidence rates before mapping," *Epidemiology,* pp. 622-630, 1994.

[20] W. Thompson, "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples," *Biometrika,* pp. 285-294, 1933.

[21] T. L. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Advances in applied mathematics,* pp. 4-22, 1985.

[22] J. Gittins, "Bandit processes and dynamic allocation indices," *Journal of the Royal Statistical Society: Series B (Methodological),* pp. 148-164, 1979.

[23] P. Auer, "sing confidence bounds for exploitation-exploration trade-offs," *Journal of Machine Learning Research,* pp. 397-422, 2002.

[24] E. Gutin and V. Farias, "Optimistic gittins indices," *Advances in Neural Information Processing Systems,* pp. 3153-3161, 2016.

[25] A. Tewari and S. A. Murphy, "From Ads to Interventions: Contextual Bandits in Mobile Health," in *Mobile Health*, SpringerLink, 2017.

[26] A. Durand, C. Achilleos, D. Iacovides, K. Strati, G. D. Mitsis and J. Pineau, "Contextual bandits for adapting treatment in a mouse model of de novo carcinogenesis," in *Machine Learning for Healthcare Conference*, 2018.

[27] S. Scott, "Multi-armed Bandit Experiments," Google Analytics, 23 January 2013. [Online]. Available: https://analytics.googleblog.com/2013/01/multi-armed-bandit-experiments.html.

[28] L. Li, W. Chu, J. Langford and R. Schapire, "A contextual-bandit approach to personalized news article recommendation," *Proceedings of the 19th international conference on World wide web,* pp. 6611-670, 2010.

[29] F. Amat, A. Chandrashekar, T. Jebara and J. Basilico, "Artwork personalization at Netflix," *Proceedings of the 12th ACM conference on recommender systems,* pp. 487-488, 2018.

[30] N. Karampatziakis, S. Kochman, J. Huang, P. Mineiro, K. Osborne and W. Chen, "Lessons from Contextual Bandit Learning in a Customer Support Bot," *arXiv preprint arXiv:1905.02219,* 2019.

[31] S. Caria, M. Kasy, S. Quinn, S. Shami and A. Teytelboym, "An Adaptive Targeted Field Experiment: Job Search Assistance for Refugees in Jordan," *CESifo Working Paper,* 2020.

[32] O. Besbes, Y. Gur and A. Zeevi, "Stochastic multi-armed-bandit problem with non-stationary rewards," *Advances in neural information processing systems,* pp. 199-207, 2014.

[33] H. Luo, C.-Y. Wei, A. Agarwal and J. Langford, "Efficient contextual bandits in non-stationary worlds," *Conference on Learning Theory,* pp. 1739-1776, 2018.

[34] H. Bastani and M. Bayati, "Online decision making with high-dimensional covariates," *Operations Research,* pp. 276-294, 2020.

[35] Z. Gao, Y. Han, Z. Ren and Z. Zhou, "Batched multi-armed bandits problem," *Advances in Neural Information Processing Systems,* pp. 503-514, 2019.

[36] V. Perchet, P. Rigollet, S. Chassang and E. Snowberg, "Batched bandit problems," *The Annals of Statistics,* pp. 660-681, 2016.

[37] S. Agrawal and N. Devanur, "Bandits with concave rewards and convex knapsacks," *Proceedings of the fifteenth ACM conference on Economics and computation,* pp. 989-1006, 2014.

[38] Hellewell, J.; Abbott, S.; Gimma, A.; Bosse, N. I.; Jarvis, C. I.; Russell, T. W.; Munday, J. D.; Kucharski, A. J.; Edmunds, W. J.; Centre for the Mathematical Modelling of Infectious Diseases COVID-19 Working Group; Funk, S.; Eggo, R. M., "Feasibility of controlling COVID-19 outbreaks by isolation of cases and contacts," *The Lancet. Global health,* 2020.

[39] W. G. Imbens and B. D. Rubin, Causal Inference in Statistics, Social and Biomedical Sciences, Cambridge University Press, 2015.

[40] P. Rosenbaum and D. Rubin, "The central role of the propensity score in observational studies for causal effects," *Biometrika,* vol. 70, no. 1, pp. 41-55, 1983.

[41] E. Y. Cramer, E. L. Ray, V. K. Lopez, J. Bracher, A. Brennen, A. J. C. Rivadeneira, A. Gerding, T. Gneiting, K. H. House, Y. Huang and others, "Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the US," *medRxiv,* 2021.

[42] A. B. Tsybakov, Introduction to Nonparametric Estimation, Springer, 2009.

[43] P. Auer, "Using confidence bounds for exploitation-exploration trade-offs," *Journal of Machine Learning Research,* pp. 397-422, 2002.

[44] H. Bastani, D. Simchi-Levi and R. Zhu, "Meta Dynamic Pricing: Transfer Learning Across Experiments," *Management Science (forthcoming),* 2021.

[45] S. Agrawal and N. Goyal, "Thompson sampling for contextual bandits with linear payoffs," *International Conference on Machine Learning,* pp. 127-135, 2013.

[46] Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics,* pp. 1189-1232, 2001.

[47] J. Baek, V. Farias, A. Georgescu, R. Levi, T. Peng, D. Sinha, J. Wilde and A. Zheng, "The limits to learning an SIR process: granular forecasting for COVID-19," *arXiv,* 2020.

[48] A. Fogg, "Anthony Goldbloom gives you the secret to winning Kaggle competitions," 13 January 2016. [Online]. Available: https://www.import.io/post/how-to-win-a-kaggle-competition/.

[49] J. Friedman, T. Hastie and R. Tibshirani, The elements of statistical learning, New York: Springer Series in Statistics, 2001.

[50] "What algorithms are most successful on Kaggle?," [Online]. Available: https://www.kaggle.com/bigfatdata/what-algorithms-are-most-successful-on-kaggle.

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.

Code Availability. The code for the deployment of the algorithm on a sample dataset is available at https://github.com/vgupta1/EvaTargetedCovid19Testing. The code for the counter-factual analysis is under preparation and will be available at https://github.com/vgupta1/Eva_CounterfactualAnalysis.

Acknowledgements. The authors would like to thank all members of the Greek COVID-19 Taskforce, the Greek Prime Minister Kyriakos Mitsotakis, the Ministry of Digital Governance, the Ministry of Civil Protection, the Ministry of Health, the National Public Health Organization, the development team from Cytech as well as the border control agents, doctors, nurses and lab personnel that contributed to Eva's deployment. Furthermore, the authors would like to thank Osbert Bastani for discussions and analysis on constructing custom risk metrics from public data. V.G. was partially supported by the National Science Foundation through NSF Grant CMMI-1661732.

Author Contributions. H.B., K.D., and V.G. constructed the model, designed and coded the algorithm, and performed the analysis in this paper. J.V. designed the software architecture and APIs to communicate with the Central Database of the Ministry of Digital Governance. C.H., P.L., G.M., D.P., and S.T. contributed to and informed epidemiological modeling choices of the system. All authors coordinated Eva's operations and logistics throughout its deployment.

Author Information. H.B., V.G., and J.V. declare no conflict of interest. K.D. declares non-financial competing interest as an unpaid Data Science and Operations Advisor to the Greek Government from May 1st, 2020 to Nov 1st, 2020. C.H., P.L., G.M., D.P., and S.T. declare non-financial competing interest as members of the Greek National COVID-19 Taskforce. Correspondence should be addressed to drakopou@marshall.usc.edu.

Appendix: Methods

1 Problem Description

Notation and System Constraints

Let $t \in \{1, \dots, \mathcal{T}\}$ index time (in days), $e \in \{1, \dots, \mathcal{E}\}$ index points of entry, and $c \in \{1, \dots, \mathcal{C}\}$ index the set of 193 countries from which travelers may originate. Moreover, for each point of entry e, let $B_e(t)$ denote the budget of available tests at time t.

Pertinent demographic data about a passenger (extracted from their PLF) include their country and region of origin[7]; since these are categorical features, we represent them with a finite, discrete set of values $\mathcal{X}$. We refer to passengers with features $x \in \mathcal{X}$ as x-passengers. Let $A_{xe}(t)$ denote the number of x-passengers arriving with date of entry t and point of entry e. For every $x \in \mathcal{X}$, let $R_x(t)$ denote the unknown, time-varying, underlying prevalence among x-passengers, i.e., the probability that a x-passenger is infected.

The budgets $B_e(t)$ were exogenously determined by the Secretary General of Public Health. It is worth noting that the availability of tests relative to arrivals varies significantly across points of entry (see Fig. 5).

Decision Problem

The main decision problem on day t is to choose the number of tests $T_{xe}(t)$ to allocate to x-passengers at point of entry e. To formally define this problem, we first specify the information available at time t. Namely, let $P_x(t)$ and $N_x(t)$ denote the number of x-passengers that tested positive and negative at time t, respectively. Then, since labs take up to two days to process testing results, the available information on day t is

$$\{P_x(t'), N_x(t')\}_{x \in \mathcal{X}, t' < t-2}.$$

Thus, at the beginning of day t, we must determine the number of tests $T_{xe}(t)$ to be performed on x-passengers at point of entry e using this information.

Furthermore, our testing decisions need to satisfy two constraints:

$$\sum_{x \in \mathcal{X}} T_{xe}(t) \leq B_e(t), \forall\, e \in \{1, \dots, \mathcal{E}\},\ \textbf{(Budget Constraint)}$$

ensuring that number of allocated tests does not exceed the budget at each point of entry, and

$$T_{xe}(t) \leq A_{xe}(t), \forall\, x \in \mathcal{X}, e \in \{1, \dots, \mathcal{E}\},\ \textbf{(Arrivals Constraint),}$$

ensuring that the number of allocated tests for x-passengers does not exceed the number of arriving x-passengers.

[7] 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.

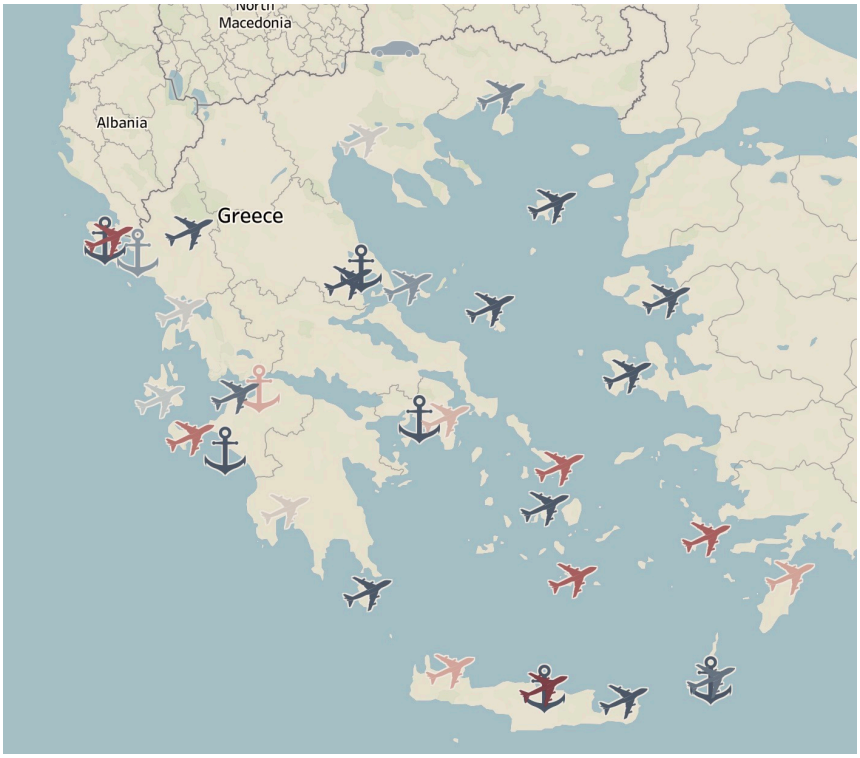


Figure 5: Selected points of entry. Airplane icons denote airports; anchor icons denote seaports; car icons denote land borders. Icon colors indicate daily availability of tests relative to daily arrival rate: blue (red) denotes relatively high (low) test supply. Note that information on some points of entry is omitted due to the sensitive nature of the data.

A secondary decision problem is to choose "color designations" as described in the main text for every country. As discussed, unlike our testing decisions, color designation decisions are not entirely algorithmic. Rather, they were made in conjunction with the Greek COVID-19 taskforce, and further reviewed/authorized by the Office of the Prime Minister.[8] In particular, Eva was only used to identify candidates for grey-listing based on its estimates of the current prevalence $R_x(t)$.

In what follows, we focus on the algorithmic elements of our procedure, i.e., determining $T_{xe}(t)$ and, as part of that process, estimating $R_x(t)$.

Objective

Note that, conditional on $T_{xe}(t)$, the number of positive tests observed at entry e is binomially distributed with $T_{xe}(t)$ trials and (unknown) success probability $R_x(t)$. Our goal is to maximize the expected total number of infections caught at the border, i.e.,

$$\mathbb{E}\left[\sum_{t=1}^{\mathcal{T}}\sum_{x\in\mathcal{X}}\sum_{e=1}^{\mathcal{E}} T_{xe}(t)R_x(t)\right].$$

[8] Greece adopted the European Union's recommended black-list designations $\mathcal{B}_{t+7}$; only Greek citizens from black-listed countries were permitted to enter, and 100% of these arrivals were tested at the border. Of the remaining countries (which accounted for over 97% of arrivals), Greece designated a subset as grey-listed countries ($\mathcal{G}_{t+7}$) and the remainder as white-listed countries ($\mathcal{W}_{t+7}$). All 3 color designations affect travelers arriving at time $t+7$ onwards; this one-week lead time is needed operationally to alert passengers of (possibly) changing travel requirements.

Note that the testing decisions we make at time t affect the information that will be available in the future, and thereby the quality of our prevalence estimates $\{\hat{R}_x(t')\}_{x \in \mathcal{X}}$ for $t' > t$.

Thus, when deciding $T_{xe}(t)$, we must balance two seemingly conflicting objectives. On the one hand, a myopic decision-maker would allocate tests to the (estimated) riskiest passengers, based on their features x and our current prevalence estimates $\{\hat{R}_x(t)\}$. On the other hand, a forward-looking decision-maker would want to collect data on x-passengers for every value of x, in order to make accurate *future* assessments on which passengers are risky. This suggests allocating tests uniformly across feature realizations to develop high-quality surveillance estimates.

This tension – known as the exploration-exploitation trade-off – is well-studied in the multi-armed bandit literature [21, 20, 22]. Optimal solutions balance the need to *explore* (accurately learn the current prevalence $R_x(t)$ for all x) and to *exploit* (use the estimated prevalence $\{\hat{R}_x(t)\}_{x \in \mathcal{X}}$ to allocate tests to the riskiest passengers).

2 Proposed Solution- Algorithm Description

2.1 Overview of Algorithm

The tradeoff presented in Section 1 resembles a multi-armed bandit problem. However, our setting exhibits a number of salient features that distinguish it from the classical formulation:

1. **Non-stationarity:** The prevalence $R_x(t)$ is time-varying.
2. **Imbalanced Data:** On average, only 2 in 1000 passengers tests positive, meaning the data $\{P_x(t'), N_x(t')\}_{x \in \mathcal{X}, t' < t-2}$ is very imbalanced with mostly negative tests.
3. **High-Dimensionality:** The number of possible features $\mathcal{X}$ is very large.
4. **Batched decision-making:** All testing allocations for a day must be determined at the start of the day (batch) for each point of entry.
5. **Delayed Feedback:** Labs may take up to 48 hours to return testing results.
6. **Constraints:** Each point of entry is subject to its own testing budget and arrival mix.

Although these features have been studied in isolation, their combination poses unique challenges. We next propose a novel algorithm that addresses these features in our setting. We separate our presentation into two parts: estimation and allocation. Our estimation procedure (Section 2.2) builds on ideas from empirical Bayes and high-dimensional statistics to address the challenges of non-stationarity, imbalanced data and high-dimensionality. Our allocation procedure (Section 2.3) uses these estimates and adapts classical multi-arm bandit algorithms to address the challenges of batched decision-making, delayed feedback, and budget constraints.

Before delving into details, we note that to address the aforementioned high-dimensionality, we periodically partition the set of features $\mathcal{X}$ into *types*, denoted by K_t. On first reading the reader can take a passenger's type to be equivalent to their country of origin (i.e., ignoring more granular feature information) without much loss of meaning. However, in actuality, we distinguish particularly risky regions within a country as additional, distinct types, where these "risky regions" are identified dynamically based on recent data (see Section 2.2 for details). After constructing the set of types K_t,

we treat all x-passengers of the same type $k \in K_t$ symmetrically in our estimation and allocation procedures, i.e., our estimates satisfy $\hat{R}_{x_1}(t) = \hat{R}_{x_2}(t)$ for all features $x_1, x_2 \in k$. This reduces the dimensionality of our estimation and allocation problems from $|\mathcal{X}|$ to $|K_t|$.

2.2 Estimation

In this section, we focus on developing estimates for prevalence rates $R_k(t)$ for a given day t; we drop the time index when it is clear from context.

Recall that the (unknown) prevalence rates $R_k(t)$ are time-varying. Common strategies for estimation in time-varying settings include discarding sufficiently old data [33, 14] or exponential smoothing [32]. The kernel smoothing literature [42] suggests that discarding old data might be preferred if the rates $R_k(t)$ are particularly non-smooth in time. Based on this intuition and conversations with epidemiologists on Greece's COVID-19 taskforce, we choose to discard testing results that are more than 16 days old. As before, let

$$P_k = \sum_{t'=t-16}^{t-3} P_k(t'),$$

denote the total number of type k passengers that tested positive over the past 14 days of test results, and let

$$N_k = \sum_{t'=t-16}^{t-3} N_k(t'),$$

denote the total number of type k passengers that tested negative over the past 14 days of test results. An unbiased, and natural estimate of the prevalence for type k is

$$\hat{r}_k^{naive} = \frac{P_k}{P_k + N_k}. \quad (1)$$

However, because prevalence rates are low (on the order of 2 in a 1000), the variability of this estimator is on the same order as $R_k(t)$ for moderate values of $T_{ke}(t)$. Worse, for rare types where $T_{ke}(t)$ is necessarily small (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 [18, 19].

As a simple illustration, consider a (foolish) baseline estimator that estimates every prevalence to be zero. We compute this baseline estimator and the naïve estimator for all countries as of Sept 1, 2020. We find that the average MSE (averaged over countries) of the naïve estimator is *larger* than that of the baseline estimator (see Fig. 6 below). In fact, a conservative estimate of the error of $\hat{r}_k^{naive}$ is

$$\sqrt{\text{Excess MSE of } \hat{r}_k^{naive}} = \sqrt{.000334} = 0.018,$$

which is larger than the typical prevalence of most countries. In other words, any potential signal is *entirely* washed out by noise.

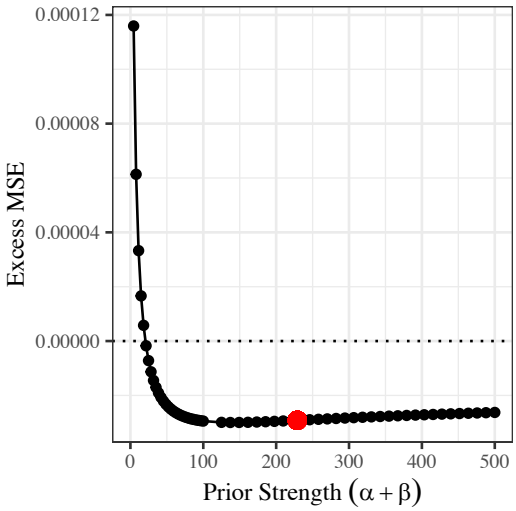


Figure 6: Excess MSE over baseline estimator (estimates zero prevalence for all types). The naïve estimator (clipped from the plot for readability) corresponds to a prior strength of zero and has far worse MSE than the baseline. Our empirical Bayes estimator (red dot) substantively improves performance and approximately minimizes MSE by trading off bias and variance.

To compensate for high variability and potentially rare types, we adopt an empirical Bayes perspective. This improves the stability of our estimates at the expense of introducing some bias. Our approach naturally allows information-sharing across types, so that rare types partially borrow data from other similar types to improve their stability.

At a high level, our estimation procedure has two steps: First, for each color designation (white-, grey- or black-listed) $l \in \{\mathcal{W}_t, \mathcal{G}_t, \mathcal{B}_t\}$, we fit a prior $\text{Beta}(\alpha_l(t), \beta_l(t))$ for all types from countries of that color-designation. Second, we compute Bayesian (posterior) estimates for each $R_k(t)$, assuming $R_k(t)$ was drawn from the appropriately colored prior.

To illustrate our prior fitting procedure, consider all countries of color $l \in \{\mathcal{W}_t, \mathcal{G}_t, \mathcal{B}_t\}$, and let $\mathcal{L}_l = \{k \in K_t : k \text{ is from a country in } l\}$. We drop the time index t for convenience. Let $T_k = P_k + N_k$ be the total number of tests (with results) allocated to type k over the last 16 days. Conditional on T_k, P_k is approximately binomially distributed with T_k trials and success probability R_k. If we further assume that for each $k \in \mathcal{L}_l$, R_k was drawn independently from a $\text{Beta}(\alpha_l, \beta_l)$ distribution, then the strong law of large numbers yields:

$$\frac{1}{|\mathcal{L}_l|}\sum_{k\in\mathcal{L}_l}\frac{P_k}{T_k} \xrightarrow{a.s.} \frac{1}{|\mathcal{L}_l|}\sum_{k\in\mathcal{L}_l}\mathbb{E}\left[\frac{P_k}{T_k} \mid T_k\right] = \frac{1}{|\mathcal{L}_l|}\sum_{k\in\mathcal{L}_l}\mathbb{E}[R_k] = \frac{\alpha_l}{\alpha_l + \beta_l}$$

and

$$\frac{1}{|\mathcal{L}_l|}\sum_{k\in\mathcal{L}_l}\frac{P_k\,(P_k-1)}{T_k(T_k-1)} \xrightarrow{a.s.} \frac{1}{|\mathcal{L}_l|}\sum_{k\in\mathcal{L}_l}\mathbb{E}\left[\frac{P_k(P_k-1)}{T_k(T_k-1)} \mid T_k\right]$$
$$= \frac{1}{|\mathcal{L}_l|}\sum_{k\in\mathcal{L}_l}\mathbb{E}[R_k^2] = \frac{\alpha_l^2}{(\alpha_l+\beta_l)^2} + \frac{\alpha_l\beta_l}{(\alpha_l+\beta_l)^2(\alpha_l+\beta_l+1)},$$

where the limits are taken as $|\mathcal{L}_l| \to \infty$. Taking the left sides of the above two expressions as estimators for the right sides, we can rearrange to find estimates of (α_l, β_l), yielding

$$\widehat{\alpha}_{\mathrm{l}} = \frac{M_{1l}^2(1-M_{1l})}{M_{2l}^2 - M_{1l}^2} - M_{1l}$$

$$\hat{\beta}_{\mathrm{l}} = \widehat{\alpha}_{\mathrm{l}} \frac{(1-M_{1l})}{M_{1l}}$$

$$M_{1l} = \frac{1}{|\mathcal{L}_l|} \sum_{k \in \mathcal{L}_l} \frac{P_k}{T_k} \tag{2}$$

$$M_{2l} = \frac{1}{|\mathcal{L}_l|} \sum_{k \in \mathcal{L}_l} \frac{P_k\,(P_k - 1)}{T_k(T_k - 1)} \tag{3}$$

We repeat this procedure separately for each of the three-color designations. Equipped with these priors, we then compute posterior distributions for each type $k \in \mathcal{L}_l$. By conjugacy, these are $\mathrm{Beta}(\alpha_{\mathrm{k}}, \beta_{\mathrm{k}})$ distributed with estimates

$$\widehat{\alpha}_k = \widehat{\alpha}_{\mathrm{l}} + \mathrm{P}_k \quad \text{and} \quad \hat{\beta}_k = \hat{\beta}_l + N_k\,,$$

suggesting the following empirical Bayes estimate of the prevalence R_k:

$$\hat{r}_k^{EB} = \frac{\widehat{\alpha}_k}{\widehat{\alpha}_k + \hat{\beta}_k}. \tag{4}$$

Below, we summarize the estimation strategy.

Estimation Strategy

Input: *# of positives P_k, negatives N_k and tests T_k for each type k in the past 14 days of test results.*

for each collection of countries $l \in \{\mathcal{W}_t, \mathcal{G}_t, \mathfrak{B}_t\}$

Compute M_{1l}, M_{2l} from Eq. (2) and (3)

$\widehat{\alpha}_{\mathrm{l}} \leftarrow \frac{M_{1l}^2(1-M_{1l})}{M_{2l}^2 - M_{1l}^2} - M_{1l}, \qquad \hat{\beta}_{\mathrm{l}} \leftarrow \widehat{\alpha}_{\mathrm{l}} \frac{(1-M_{1l})}{M_{1l}}$

for all types $k \in \mathcal{L}_l$ whose origin country is in l:

$\widehat{\alpha}_k \leftarrow \widehat{\alpha}_{\mathrm{l}} + P_k$ and $\hat{\beta}_k \leftarrow \hat{\beta}_{\mathrm{l}} + N_k$

end

end

To provide intuition, notice that if type k comes from a country with color designation l, our posterior estimate can be rewritten as

$$\hat{r}_k^{EB} = \left(1 - \frac{T_k}{T_k + \alpha_{\mathrm{l}} + \beta_{\mathrm{l}}}\right) \frac{\alpha_{\mathrm{l}}}{\alpha_{\mathrm{l}} + \beta_{\mathrm{l}}} + \frac{T_k}{T_k + \alpha_{\mathrm{l}} + \beta_{\mathrm{l}}} \hat{r}_k^{naive},$$

i.e., it is a weighted average between the naïve prevalence estimator and our prior mean. The sum $(\alpha_{\mathrm{l}} + \beta_{\mathrm{l}})$ is often called the strength of the prior. For rare types where T_k is small relative to $(\alpha_{\mathrm{l}} + \beta_{\mathrm{l}})$, the estimator is close to the prior mean (it draws information from similar types). For common types where T_k is large relative to $(\alpha_{\mathrm{l}} + \beta_{\mathrm{l}})$, it is close to the naïve estimator, matching our intuition that the naive estimator should only be used T_k is large enough.

Fig. 6 shows the excess MSE for various estimators with differing strength in the prior (the naïve estimator $\hat{r}_k^{naive}$ corresponds to a prior strength of zero and its MSE is too large to fit on the plot). Note that our moment-matching estimator (red dot) approximately minimizes the MSE while maintaining a tractable closed-form expression.

Reducing Dimensionality through Adaptive Type Identification

As mentioned earlier, we partition the discrete space of features $\mathcal{X}$ into a smaller set of types K_t to reduce the dimensionality of the problem. Defining types entirely by a passenger's country of origin is attractive for its simplicity and because geography is highly predictive of prevalence. That said, there can be significant heterogeneity in infection prevalence within a country. For example, certain regions may have a high population density and low social distancing compliance relative to the rest of the country, resulting in a much higher risk for passengers originating from that particular region. Defining types at the country-level risks poor test allocations by failing to test passengers from risky regions within otherwise safe countries.
Consequently, our dimensionality reduction procedure starts with types defined at the country level, but then goes one step further to distinguish particularly risky regions[9] within a country as additional, distinct types. These "risky regions" are identified dynamically based on recent testing results. 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 [17], which has also previously been used for dimensionality reduction in contextual bandits [34]. Specifically, we first apply our previous empirical Bayes estimation strategy assuming country-based types. Given these estimates, we then perform a LASSO logistic regression on all testing results within the last 14 days of testing results, where the unit of observation is a single passenger's test, and features include (1) the estimated prevalence $\hat{r}_c^{EB}$ of the passenger's origin country (estimated in previous step) and (2) dummy variables for potential regions.

Mathematically, let $\boldsymbol{y}$ denote the vector of $\{0,1\}$ test results over the past 14 days of available results, let f_i index potential regions, and c_i index potential countries. We perform the LASSO logistic regression:

$$y_i = \sum_{c=1}^{C} \delta_c \mathbf{1}(c = c_i)\hat{r}_c^{EB} + \sum_f \delta_f \mathbf{1}(f = f_i) + \epsilon_i.$$

This yields a sparse vector of coefficients $\left[\hat{\delta}_c, \hat{\delta}_f\right]$. The nonzero support of $\hat{\delta}_f$ provides a set of regions whose prevalence is notably different than the prevalence of the corresponding country based on recent testing results. Note, for our purposes, it is only useful to determine significantly *riskier* regions (so we can focus our limited testing resources on passengers from those regions), which corresponds to achieving a positive coefficient in $\hat{\delta}_f$. Thus, we define new types for any regions f_i that satisfy $\left(\hat{\delta}_f\right)_i > 0$.

The process is summarized below:

[9] Regions indicates more granular locations within a country (e.g., state or province). Since different countries use different terms, we use the generic term "region."

Adaptive Definition of Types

Input: y_i, c_i, f_i *# historical test results with country and region dummies*

Perform Lasso regression on $y_i = \sum_{c=1}^{C} \delta_c \mathbf{1}(c = c_i)\hat{r}_c^{EB} + \sum_f \delta_f \mathbf{1}(f = f_i) + \epsilon_i$

return set of types $K_t = \{1, \dots, C\} \cup \left\{f_i : \left(\hat{\delta}_f\right)_i > 0\right\}$

We re-ran this procedure every week to obtain new types K_t, which are used daily for empirical Bayes estimation of prevalence rates as well as the bandit allocation algorithm.

2.3 Allocating Test Results

Optimistic Gittins Indices

Online optimization problems and the aforementioned exploration-exploitation tradeoff have been studied in the literature since the seminal work of [20, 21]. In the classical setting, the celebrated Gittins index theorem provides the optimal dynamic solution to this tradeoff [22], but it is typically intractable to compute this solution exactly. Consequently, various heuristics with near-optimal asymptotic performance guarantees such as Upper Confidence Bound [43], Thompson Sampling [20] and optimistic Gittins indices [24] have been proposed to overcome these issues. Our approach builds upon the optimistic Gittins index of [24]. We first describe the mechanics of the optimistic Gittins index in a classical setting, and then describe how we adapt this technique to address the unique combination of challenges in our setting – batched decision-making, delayed feedback and port-specific budget/arrival constraints.

Recall that our estimation procedure yields a Beta posterior distribution with parameters α_k, β_k for type k. Let $F_{\alpha,\beta}$ be the CDF of a Beta distribution with parameters α and β. Then, from [24], the Optimistic Gittins Index λ_k for type k (with a 1-step lookahead window and a discount factor γ) is the unique solution to the following equation:

$$\lambda_k = \frac{\alpha_k}{\alpha_k + \beta_k}\left(1 - \gamma F_{\alpha_k+1,\beta_k}(\lambda_k)\right) + \gamma\lambda_k\left(1 - F_{\alpha_k,\beta_k}(\lambda_k)\right). \quad (5)$$

In the classical setting, the Optimistic Gittins Index Algorithm proceeds very simply: (1) test the type (arm) with the highest index, (2) observe the resulting (immediate) feedback and use it to update the posterior for that type, (3) calculate the new index for that type, and iv) repeat. Importantly, 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 $\hat{r}_k^{EB}$).

Challenges with the Conventional Optimistic Gittins Index Algorithm

Unfortunately, this approach does not perform well in our setting because we do not observe *immediate* feedback from our tests. Rather, we choose thousands of passengers to test in a given day (avg: 5300; std dev: 998) and receive no feedback on these allocations for 48 hours. Applied naively, the algorithm would simply keep testing the same type (the one with the highest initial index) repeatedly because the computed indices for each type remain the same throughout a given batch (day). Such an outcome is clearly undesirable.

The batched bandit literature [36, 35] partially resolves this issue in stationary environments by uniformly exploring all types in early batches, and then committing to the type with the highest prevalence in later exploitation batches. However, this strategy is untenable in highly non-stationary environments, because the data from initial exploration in early batches are not representative of current rates (recall that we only use data from the last 16 days for estimation), i.e., we must continuously explore and collect new data on each type to form accurate prevalence estimates. Thus, it is critical in our setting that our allocation policy *combine* exploration and exploitation *within* a batch.

Relatedly, at the start of each batch, we also have a large number of tests that are already in the "pipeline" (i.e., tests that have been conducted but the results have not yet been received from the labs) from the past 2 days. A good policy should also account for the expected information that will be imbued by these pipeline tests when making new allocations.

Priority Pseudo-Updating

In order to address the delayed-feedback and batching challenges above, we propose **certainty-equivalent** pseudo-updates to our indices during the course of the allocation assignment algorithm. These updates do not alter the mean estimates of prevalence but reduce the variance of our estimates based on the number of tests allocated thus far, thereby anticipating information that we will obtain on uncertain types in the next few days.

The intuition is as follows: Although we do not observe immediate feedback when allocating a test, we can estimate the likely reduction in the variance of our posterior distributions. Specifically, if we ignore the uncertainty (motivating our nomenclature certainty-equivalent) around our estimate $\hat{r}_k^{EB}$, then, after allocating a single additional test to type k, we expect to observe a positive result with probability $\hat{r}_k^{EB}$ and a negative result with probability $1 - \hat{r}_k^{EB}$ in 48 hours. Consequently, after allocating a test, we update the parameters of our Beta distribution with this expected result, namely: $\hat{\alpha}_k \leftarrow \hat{\alpha}_k + \hat{r}_k^{EB}$, and $\hat{\beta}_k \leftarrow \hat{\beta}_k + 1 - \hat{r}_k^{EB}$. This update does not change our estimate of the mean prevalence, but it does reduce the variance of our posterior distributions. In this sense, it quantifies the additional information we are likely to get after the delayed feedback. We refer to this procedure as a *pseudo-update* to the optimistic Gittins index, since we are performing a certainty-equivalent update based on an allocated test whose feedback has yet to be observed (in contrast to a Bayesian update based on feedback received from an allocated test). Importantly, pseudo-updates allow the optimistic Gittins indices to dynamically change during the course of allocation assignment within a batch.

Overall, these pseudo-updates ensure that our algorithm allocates a *minimum* number of tests required to resolve uncertainty for types with high variance (exploration) and allocates all remaining tests to arms with high estimated prevalence (exploitation). The pseudocode is presented below.

Gittins Pseudo-Update for type k

Input: $\hat{\alpha}_k, \hat{\beta}_k$ for type k

$\hat{r}_k^{EB} \leftarrow \dfrac{\hat{\alpha}_k}{\hat{\alpha}_k + \hat{\beta}_k}$

$\hat{\alpha}_k \leftarrow \hat{\beta}_k + \hat{r}_k^{EB}$ and $\hat{\beta}_k \leftarrow \hat{\beta}_k + 1 - \hat{r}_k^{EB}$

Compute λ_k from Eq. (5) using $\hat{\alpha}_k, \hat{\beta}_k$

return λ_k

Prior Widening

Unlike the theoretical formulation in [22, 24], our empirical Bayesian priors are data-driven and may therefore be *mis-specified* (e.g., due to our model assumptions, estimation error or rapid changes in prevalence that have not yet been reflected in recent testing results). The Bayesian bandit literature has shown that "widening" posterior distributions (i.e., inflating variance) can ensure that the resulting allocations are robust to prior misspecification [44, 45]. Accordingly, at the start of each batch, we decrease the strength of our prior by scaling down our estimated posterior parameters for each type k by a constant $c \in (0,1)$:

$$\hat{\alpha}_k \leftarrow c\hat{\alpha}_k, \qquad \hat{\beta}_k \leftarrow c\hat{\beta}_k.$$

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. Typically, $c \in [.1, .5]$.

Test Allocation Strategy

Equipped with our pseudo-update technique, we can now describe the allocation procedure that is run each day. Let Q_k denote the number of pipeline tests for type k (i.e., tests allocated to type k passengers in the last 2 days for which we have yet to receive feedback).

If we had a single point of entry, we would proceed by first estimating the posterior distributions of each type using our empirical Bayes strategy, and widening the resulting posterior parameters. We would then perform Q_k pseudo-updates for type k to account for the expected information imbued by current pipeline tests. We would then allocate a test to the type with the highest optimistic Gittins index for which there are still available untested passengers, perform a pseudo-update to the posterior for that type, and repeat until we deplete our testing budget or run out of passengers. Notice this entire procedure can be run at the beginning of the day, and, hence, the allocations are pre-computed.

However, since we have multiple ports of entry, we must also account for constraints on port-specific testing budgets and arrivals. Specifically, once we have identified that type k has the highest (current) optimistic Gittins index, we must decide to *which* type k passenger we will allocate a test. This decision is not entirely trivial. Depending on the choice, it will consume testing budget at a particular point of entry. Some points of entry have very limited testing budgets (see Fig. 5), and passengers of some types only travel to certain points of entry. For example, as shown in Fig. 7 below, passengers from North Macedonia (MK), a rare type, only arrive at a few points of entry, while passengers from Germany (DE), a common type, arrive across many points of entry. Intuitively, one should not allocate tests to common passenger types at a port of entry predominantly used by rare types. Rather, one should strategically "save tests" at that point of entry for potential rare type arrivals. Indeed, if we exhaust the budget on common passenger types early in the batch, we will be unable to test rare passenger types if directed to by the Gittins procedure later on in the batch.

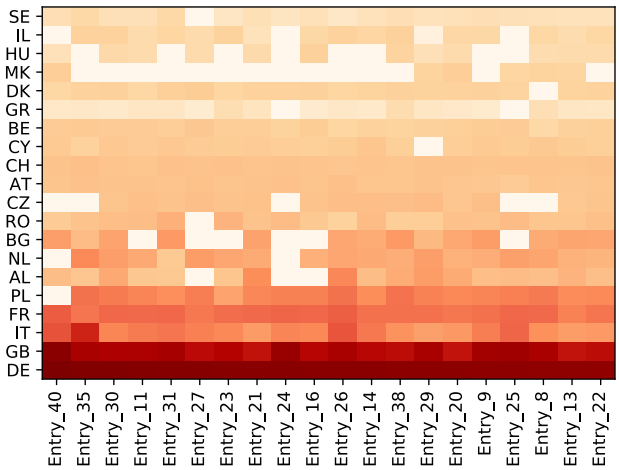


Figure 7 Heatmap of arrivals from selected countries (y-axis) to selected ports (x-axis). Darker color signifies more arrivals.

Although the bandit literature has explored incorporating constraints [37] into allocations, it remains an open problem to adapt such approaches to a nonstationary, batched environment (e.g., incorporating constraints along with our certainty-equivalent updates). An optimal allocation can be determined by solving a large binary integer program, but solution times can be long.[10] Instead, we employ a greedy heuristic to resolve this last challenge. Specifically, if type k currently has the highest optimistic Gittins index, we choose a passenger of type k at the point of entry with the most remaining tests available. In this manner, we preferentially allocate tests at less constrained ports, with the goal of potentially saving tests for rare types at constrained ports. When a point of entry's budget is depleted, we remove that point of entry from consideration; similarly, if all arrivals of a certain type have been assigned to be tested, we remove that type from consideration. Otherwise, we follow the procedure outlined earlier. The pseudocode for our test allocations is provided below. Once we obtain the allocations, we randomly select the requisite number of passengers of each type at each point of entry.

[10] 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.

Test Allocation Sub-Routine

Input: Posterior distribution estimates $\{\hat{\alpha}_k, \hat{\beta}_k\}_{k \in K_t}$ for each type k, arrivals $A_{ke}(t)$ and budgets $B_e(t)$ for each type k and point of entry e, number of pipeline tests Q_k, tuning parameters $\{c, \gamma\}$

$B(t) \leftarrow \sum_e B_e(t)$, $A_k(t) \leftarrow \sum_e A_{ke}(t)$ *# total testing budget and total type k arrivals*

for $k = 1{:}\,K_t$

$\hat{\alpha}_k \leftarrow \mathrm{c}\,\hat{\alpha}_k, \hat{\beta}_k \leftarrow \mathrm{c}\,\hat{\beta}_k$ *# prior widening*

Compute λ_k from Eq. (5) using $\hat{\alpha}_k, \hat{\beta}_k$

$\lambda_k \leftarrow$ pseudo-update index of type k (repeat Q_k times) *# account for pipeline tests*

$Y_k \leftarrow A_k(t)$ *# type k passengers not yet allocated a test*

for $e = 1{:}\,\mathcal{E}$

$Y_{ke} \leftarrow A_{ke}(t)$ *# type k passengers at point of entry e not yet allocated a test*

$C_e \leftarrow B_e(t)$ *# remaining (un-allocated) tests at point of entry e*

$N_{ke} \leftarrow 0$ *# initialize allocations*

end

end

while $\max_{\mathrm{e}} C_e > 0$ **and** $\max_{\mathrm{k},\, e' \in \{e \mid C_e > 0\}} Y_{ke'} > 0$

$k^* = argmax\{\lambda_k : Y_k > 0\}$

$e^* = argmax\{C_e : Y_{k^* e} > 0\}$

$N_{k^* e^*} \leftarrow N_{k^* e^*} + 1$ *# allocate a test to a type k passenger at point of entry e*

$C_{e^*} \leftarrow C_{e^*} - 1,\ Y_{k^* e^*} \leftarrow Y_{k^* e^*} - 1,\ Y_{k^*} \leftarrow Y_{k^*} - 1$

$\lambda_{k^*} \leftarrow$ pseudo-update index of type k^*

end

return $\{N_{ke}\}_{k \in K_t, e \in \{1,\ldots,\mathcal{E}\}}$ *# number of tests allocated to each type at each port of entry*

Fig. 8 shows the resulting (anonymized) allocations for a typical day (batch) from the summer of 2020. The x-axis indexes different types ordered from highest prevalence (left) to lowest prevalence (right). Each bar denotes the number of arrivals. The teal portion represents the number of allocated tests, while the pink portion represents the remaining untested passengers. We observe that, as desired, our algorithm allocates tests to essentially all high-risk arrivals (exploitation) but additionally assigns a small number of tests to all types to reduce the variance of our estimates (exploration). The inclusion

of port-specific constraints may cause some distortions, e.g., the lowest-risk type in Fig. 8 still receives a large number of tests because these passengers arrived at a point of entry with atypically large testing capacity.

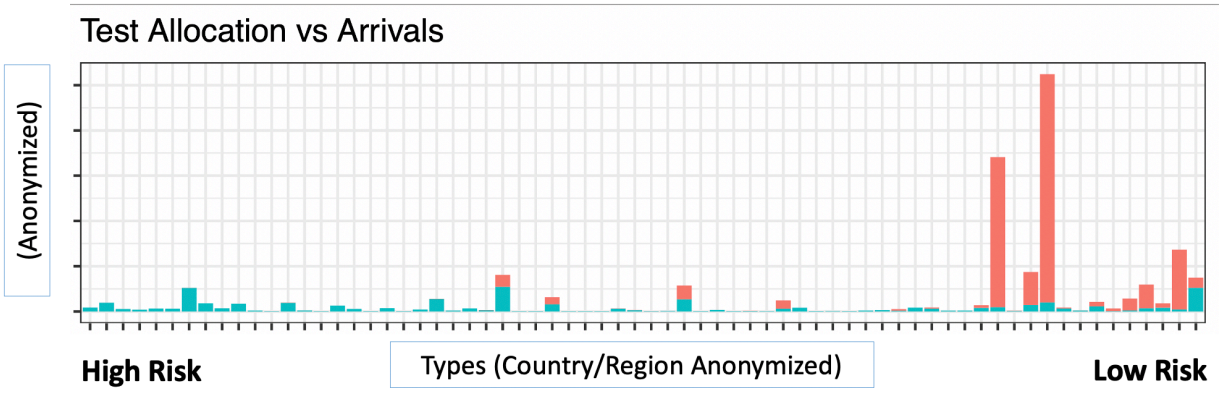


Figure 8: Actual number of test allocations (teal) and scheduled arrivals (pink) for each type on a given day (batch). Types are ordered from high to low risk according to our empirical Bayes estimates of prevalence. Note that our algorithm allocates tests to essentially all high-risk arrivals (exploitation) but additionally assigns a small number of tests to all types (exploration).

3 Off-Policy Evaluation and Counterfactual Analysis

In this section, we detail our comparison of Eva's historical performance to a natural and popular baseline: random surveillance testing. Random surveillance testing is not data-driven and therefore, does not require any of Eva's operational infrastructure.

Specifically, we benchmark against a random surveillance testing policy that tests passengers uniformly at random at each port of entry and has access to the same testing budgets as Eva did. In our initial analysis in Section 3.1, we assume that this policy also makes the same grey-listing decisions as Eva did. This assumption is particularly optimistic for the benchmark, because in the absence of Eva's operational infrastructure, grey-listing decisions would necessarily be based on publicly available epidemiological data. As we show in Section 4, such public data offers limited predictive value for the actual prevalence of asymptomatic travelers, and further carries an information delay of approximately 9 days. Thus, in Section 3.3, we provide an additional analysis comparing Eva to random surveillance testing policy that made grey-listing decisions with a 9-day lag relative to Eva's decisions.

Before proceeding, we note that the historically realized data from Eva exhibits a significant no-show rate, i.e., passengers who filed a PLF but did not travel. 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. Denote by $\hat{T}_{ke}(t)$ the number of passengers of type k that were actually tested at point of entry e on day t (as acknowledged by scanning their QR-code and associating it with a sample). Due to no-shows, this number may be less than $T_{ke}(t)$, the number of passengers of type k that filed a PLF and were allocated a test by Eva. We estimate the type-specific fraction of passengers that actually arrived by

$$s_{ke}(t) = \frac{\hat{T}_{ke}(t)}{T_{ke}(t)},$$

and estimate the actual arrivals by $\hat{A}_{ke}(t) = s_{ke}(t)\, A_{ke}(t)$. This estimate is unbiased because our decision to test a traveler is conditionally independent of their decision to not travel ("no-show") given their type.

3.1 Comparison to Random Surveillance Testing

While we directly observe Eva's performance, we do not observe the performance of random testing, requiring us to estimate a *counterfactual*. We use a classical method, inverse propensity weight (IPW) scoring [39, 40], which is a model-agnostic approach. In particular, although our estimation and allocations entail several approximations, the IPW estimate of performance is conditionally unbiased (conditional on the arrival process of passengers) *regardless of the quality of these approximations*. Thus, this analysis provides a fair comparison between the two strategies. We perform all analysis conditional on the number and types of passengers that arrive at each time at each point of entry.

We next summarize the pertinent details of the IPW method. Recall that $P_{ke}(t)$ denotes the total number of positive type k passengers identified by Eva. Let

- $\hat{T}_{\cdot e}(t) = \sum_{k \in K_t} \hat{T}_{ke}(t)$ denote the total number of tests performed at entry e
- $\hat{T}_{k\cdot}(t) = \sum_{e=1}^{\mathcal{E}} \hat{T}_{ke}(t)$ denote the total number of tests performed on type k passengers
- $\hat{A}_{\cdot e}(t) = \sum_{k \in K_t} \hat{A}_{ke}(t)$ denote the (estimated) actual arrivals at entry e
- $\hat{A}_{k\cdot}(t) = \sum_{e=1}^{\mathcal{E}} \hat{A}_{ke}(t)$ denote the (estimated) actual type k arrivals at any entry.

Under random surveillance testing, the probability a type k arrival would be tested at time t is

$$\sum_{e=1}^{\mathcal{E}} \frac{\hat{T}_{\cdot e}(t)}{\hat{A}_{\cdot e}(t)} \frac{\hat{A}_{ke}(t)}{\hat{A}_{k\cdot}(t)}.$$

By contrast, the probability a type k arrival would be tested at time t by Eva is

$$\sum_{e=1}^{\mathcal{E}} \frac{\hat{T}_{ke}(t)}{\hat{A}_{ke}(t)} \frac{\hat{A}_{ke}(t)}{\hat{A}_{k\cdot}(t)} = \frac{\hat{T}_{k\cdot}(t)}{\hat{A}_{k\cdot}(t)}.$$

Recall that $P_{k\cdot}(t)$ is the number of positive type-k cases identified at time t by Eva. IPW scoring estimates the corresponding number for random testing by

$$f_k(t) = \left(\sum_{e=1}^{\mathcal{E}} \frac{\hat{T}_{\cdot e}(t)}{\hat{A}_{\cdot e}(t)} \frac{\hat{A}_{ke}(t)}{\hat{A}_{k\cdot}(t)} \right) \Bigg/ \left(\frac{\hat{T}_{k\cdot}(t)}{\hat{A}_{k\cdot}(t)} \right) \times P_k(t),$$

where the leading ratio is called the inverse propensity weight. In words, IPW scoring corrects the observed number of positives by multiplying by the relative likelihood of being tested under both methods. We then estimate the total number of positive cases that random surveillance would have caught by summing:

$$I^{RS} = \sum_{t=1}^{\mathcal{T}} \sum_{k \in K_t} f_k(t)$$

It is common practice in the causal inference literature to drop elements of the summand for extreme values of the inverse propensity weight [39] to reduce variability. We follow this practice and drop any outlier (e, t, k) combinations where the inverse propensity score exceeds the 97.5% quantile (i.e.,

larger than 34.7). To ensure fair comparison, we also drop the same elements from Eva's performance when reporting relative improvements.

Estimating the Variance

We now turn to estimating the variance of I^{RS}. Although I^{RS} is conditionally unbiased in a model-agnostic fashion, we require some modeling assumptions to approximate its variance. Specifically, we compute the conditional variance of I^{RS} 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 on the estimated prevalence, which depends on the infection status of individuals in the last 14 days of testing results. However, for the following reasons, we expect this dependence to be quite small:

- Our prior-widening heuristic forces a certain minimal amount of exploration of each type (roughly 500 tests every 14 days), which occurs regardless of the results of previous tests. Thus, a large number of "exploration tests" are allocated independently of past test results.
- By design, our allocation procedure only depends on the estimated prevalence, not on the infection status of individual passengers. Thus, if we test a large number of passengers of a particular type, our estimated prevalence (a random variable) will be very close to the true prevalence (an unknown constant, and therefore independent of prior testing results). The same argument follows for allocation procedures based on these prevalence estimates.
- Finally, for rare types for which we may not have accurate estimates, our allocation procedure tests 100% of the arrivals regardless of past testing results, again inducing independence. (This feature is a consequence of our allocation procedure; types with diffuse priors have high optimistic Gittins indices.)

Under this assumption, the estimates $f_k(t)$ above are conditionally independent (given the arrival process) and distributed as (scaled) binomial random variables. We adapt the standard estimator of variance of a binomial distribution, yielding,

$$\begin{aligned} Var\big(f_k(t)\big) &\approx \left(\sum_{e=1}^{\mathcal{E}} \frac{\hat{T}_{\cdot e}(t)}{\hat{A}_{\cdot e}(t)} \frac{\hat{A}_{ke}(t)}{\hat{A}_{k\cdot}(t)}\right)^2 \Bigg/ \left(\frac{\hat{T}_{k\cdot}(t)}{\hat{A}_{k\cdot}(t)}\right)^2 \times Var(P_{k\cdot}(t)) \\ &\approx \left(\sum_{e=1}^{\mathcal{E}} \frac{\hat{T}_{\cdot e}(t)}{\hat{A}_{\cdot e}(t)} \frac{\hat{A}_{ke}(t)}{\hat{A}_{k\cdot}(t)}\right)^2 \Bigg/ \left(\frac{\hat{T}_{k\cdot}(t)}{\hat{A}_{k\cdot}(t)}\right)^2 P_{k\cdot}(t)\left(1 - \frac{P_{k\cdot}(t)}{\hat{T}_{k\cdot}(t)}\right). \end{aligned}$$

We then approximate $Var(I^{RS}) \approx \sum_{t=1}^{\mathcal{T}} \sum_{k \in K_t} Var\big(f_k(t)\big)$.

3.2 Grey-list Counterfactuals

Grey-listing has two effects: it reduces the prevalence of infections among arrivals (because all arrivals are required to obtain a negative PCR test) and it reduces arrivals (because it is difficult or inconvenient for tourists to obtain a test). In order to assess the impact of our grey-listing policy,

we must estimate the counterfactual prevalence and arrival rate that would have occurred had we not grey-listed a nation. For both estimates, we use fitting gradient boosted regression models[11] [46] and tune parameters by 5-fold cross-validation. Specifically, our models are as follows:

Counterfactual Prevalence Model: The unit of observation is a country-date pair and the response is the prevalence estimated by Eva. We have training data from white-listed and grey-listed countries *prior* to grey-listing, as well as training data from white-listed countries *post* grey-listing. Our features included time series data of publicly reported cases, deaths and testing rates for each day within a $[-20, +20]$ day range, a categorical country variable to control for fixed-effects, as well as the date (to capture global time trends). Note that we include data from future dates to account for the information lag in public data.

The resulting prevalence model was used to predict the prevalence for grey-listed countries prior to grey-listing (in-sample) and up to 9 days *post* grey-listing (out-of-sample); the latter yields counterfactual prevalence estimates to infer the number of infections prevented by grey-listing. The left panel of Fig. 4 in the main body depicts the results for Malta; as desired, we observe a close match between the true and counterfactual prevalence prior to grey-listing.

Observe that our prevalence estimates indicate a sustained drop in prevalence among arrivals from the grey-listed country. In particular, one week after grey-listing, we still see a 44% reduction on average in prevalence for grey-listed countries relative to the non-grey-listed counter factual estimates.

Counterfactual Arrivals Model: The unit of observation is every grey-listed country-date pair and the response is the 7-day rolling average number of arrivals from that country on that day. We have training data from all countries *prior* to grey-listing, as well as training data from non-grey-listed countries *post* grey-listing. Our features included current total arrivals from white-listed countries, total arrivals from black-listed countries, a categorical country variable, as well as the date (to capture global time trends). One concern is that, as discussed earlier, a significant fraction of scheduled arrivals does not materialize at the border, and this varies by country. However, our data suggests that the no-show rate does not materially change due to the grey-listing policy. Thus, we compute a single no-show rate per country, and use this as a multiplier on both actual and predicted arrival rates to calculate actual arrivals.

The resulting arrivals model was used to predict arrivals for grey-listed countries prior to grey-listing (in-sample) and up to 9 days *post* grey-listing (out-of-sample); the latter yields counterfactual arrival estimates to infer the number of infections prevented by grey-listing. The right panel of Fig. 4 in the main body depicts the results for Malta; as desired, we observe a close match between the true and counterfactual arrivals prior to grey-listing.

These predictions also indicate a sustained drop in arrivals from the grey-listed country. In particular, one week after grey-listing, we see a 39% reduction on average in arrivals for grey-listed countries relative to the non-grey-listed counter factual estimates.

[11] We used the R implementation publicly available at: https://cran.r-project.org/web/packages/gbm/index.html

3.3 From Counterfactual Analysis to the Value of Grey-listing

We use our above counterfactual analysis to estimate the value of grey-listing. Specifically, for each grey-listed country and for each day in the 9-day window where Eva would have grey-listed but the baseline policy did not, we compute

$$\hat{A}_k^{GCF}(t)\,\hat{r}_k^{GCF}(t) - \hat{A}_k(t)\,\hat{r}_k^{EB}(t),$$

where $\hat{r}_k^{GCF}(t), \hat{A}_k^{GCF}(t)$ are the model predictions for prevalence and arrivals under the grey-listing counterfactual from the previous subsection. The above difference is then an estimate of the total number of infected arrivals in the non-grey-listed scenario minus the total number of infected arrivals in the grey-listed scenario. Notice all these infections were prevented by Eva since none of them arrived in Greece (they remained home and did not travel). We sum this contribution up for every day in the 9 days following grey-listing for every country that was grey-listed to form our estimate of the value of grey-listing in the main paper.

Estimating the Variance

We next estimate the variance of this value of grey-listing conditional on the arrival process. The dominant term in the variance stems from the variability of $\hat{A}_k^{GCF}(t)\,\hat{r}_k^{GCF}(t)$, and, in particular, from variability in our estimation procedure from Section 3.2. Hence, we focus on quantifying this variability.

Specifically, for each predictive model from Section 3.2, we evaluate the prediction error by splitting our original training dataset into a training set (70% of observations) and test set (30% of observations). The residuals on the test set $y - \hat{y}$ appear roughly normally distributed in both cases. Denote the variance of the residuals as σ_{prev}^2 for predicting prevalence, and $\sigma_{arr}^2(c)$ for predicting arrivals for country c.[12] Thus, we model our counterfactual estimates of prevalence and arrivals as normal random variables, centered around the predicted value with variance given by σ_{prev}^2 and $\sigma_{arr}^2(c)$ respectively. We then compute the value of grey-listing 9 days later than Eva (as described above) across 1 million Monte Carlo simulations to numerically estimate its variance.

4 Evaluation of Public Data

We now turn to publicly reported data on cases, deaths and testing rates for countries around the world [10, 11, 12]. Recently, [47] used mathematical modeling to show that forecasting within epidemiological SIR models (through data-driven parametrization of model parameters) is not effective in predicting spikes in infections.

4.1 Predictive Value of Commonly Used Epidemiological Data

In order to evaluate whether public data can substitute for private data collected at the Greek border, we retrospectively used public data to predict whether or not a country was high-risk (and therefore a candidate for grey-listing and/or frequent testing at the border). Since a cutoff of 0.5% was typical for

[12] We used a single model to predict prevalence for all countries, but we used a separate country-specific model to predict arrivals for each grey-listed country. These decisions were made to improve out-of-sample accuracy.

initiating grey-listing discussions with the Greek COVID-19 taskforce, we define a country as high-risk if it has more than 0.5% prevalence (as estimated by Eva) and low-risk otherwise. However, our results are qualitatively similar across a range of cutoffs.

We predict this binary response variable using gradient boosted regression models [46] with 500 trees and 5-fold cross-validation. These models have been shown to perform well across a variety of prediction tasks [48], but we also tested other predictive models such as LASSO and recurrent neural networks (RNNs) and obtained similar results.

We evaluated different combination and granularities of public data, presented below:

- Model 1: Features are the 14-day average of cases and deaths per capita.
- Model 2: Features are the 14-day average of cases, deaths and tests performed per capita, and the positivity rate (number of positives divided by number of tests).
- Model 3: Features are the 14-day timeseries of cases and deaths per capita.
- Model 4: Features are the 14-day timeseries of cases, deaths and tests performed per capita, and the positivity rate.
- Model 5: Features are the 14-day timeseries of cases, deaths and tests performed per capita, the positivity rate, and country fixed effects.

For each model, we use the predictive performance on a held-out test set to assess the performance of our different estimators. Performance is measured by AUROC (area under ROC curve), which is more reliable than accuracy in imbalanced data [49].

4.2 Evaluation of Delay Between Prevalence and Public Case Data

For each country, we aim to classify whether it is currently at risk, i.e., its prevalence (as measured by Eva) exceeds its median prevalence over the summer. Intuitively, if a country's case data lags its underlying prevalence by ℓ days, then we can much more effectively predict its risk status y_t on day t using case data from ℓ days *in the future*, i.e., cases in the period $[t + \ell - 14, t + \ell]$. For each country, we build separate gradient boosted regression models for each ℓ ranging from 1 to 20. The left panel of Fig. 9 shows the resulting AUROC as a function of ℓ for 3 representative countries. A large AUROC for some ℓ suggests an information delay of ℓ days, e.g., France exhibits a delay of 8-9 days.

We then use mixed-integer optimization to group countries into a small number of clusters with similar delays ℓ. In particular, we take as input the AUROC $E_{c\ell}$ of the model for country c with lag ℓ based on the above analysis. Then, for a given value of M clusters, we solve the following binary optimization problem:

$$\max_{z,y} \sum_{c=1}^{\mathcal{C}} \sum_{\ell=1}^{20} E_{c\ell} z_{c\ell}$$
$$s.t.\ z_{c\ell} \in \{0,1\}$$
$$z_{c\ell} \leq y_\ell \text{ for all } c \in \{1, \dots, \mathcal{C}\}$$
$$\sum_{\ell=1}^{20} z_{c\ell} = 1 \text{ for all } c \in \{1, \dots, \mathcal{C}\}$$
$$\sum_{\ell=1}^{20} y_\ell \leq M$$

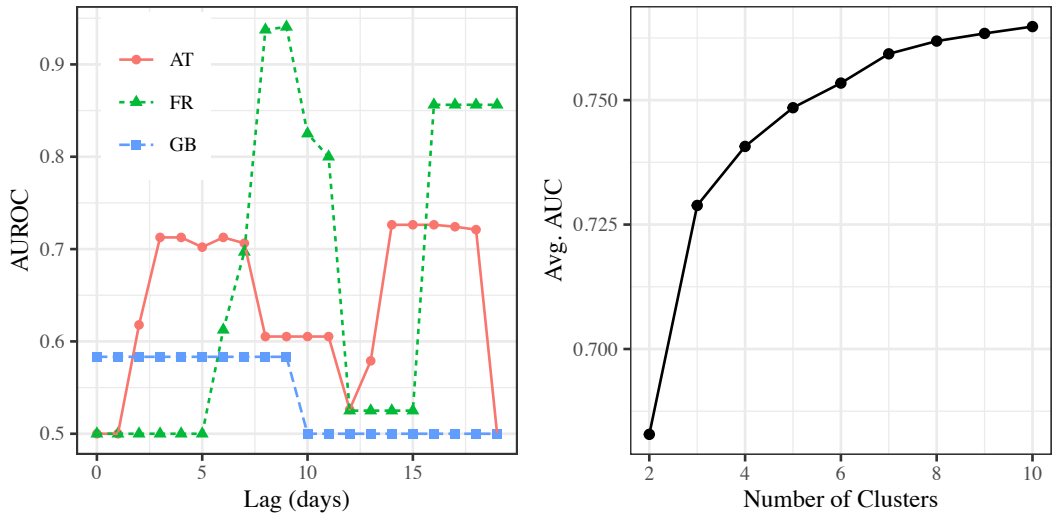
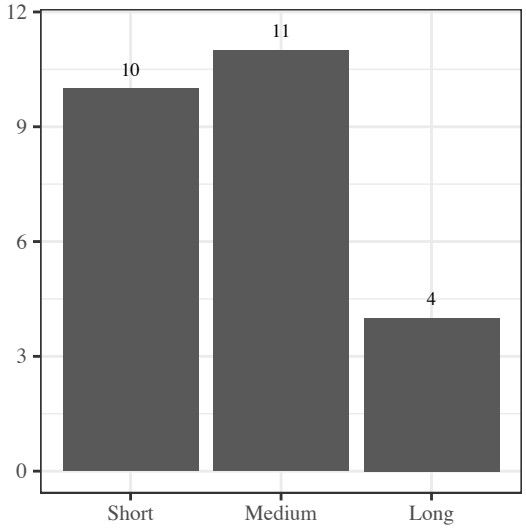


Figure 9. Left: AUROC for various ℓ for three representative countries: Austria (AT), France (FR), and Great Britain (GB). Large AUROC at lag ℓ suggests an information delay of ℓ days. Right: Elbow graph of the objective value of the optimization problem vs the number of clusters M. The sharp change in slope at 3 strongly supports three clusters in the data.

This optimization assigns each country to a single lag, encoded by the decision variable $z_{c\ell}$. Every country must be assigned some lag (the first equality constraint). The first inequality constraint ensures that the variables y_ℓ will be equal to 1 and non-zero only if some country was assigned lag ℓ. The final inequality constraint ensures that the number of distinct lags to which countries are assigned is no more than M. In other words, the optimization seeks a clustering of the countries to lag with minimal error and no more than M clusters. We solved this optimization for various values of M and used the common heuristic of looking for an "elbow" in the resulting objective values (see left panel of Fig. 9) to select $M = 3$.

Fig. 10 identifies the resulting three clusters of delays: short (1 day), medium (9 days) and long (16 days), as well as representative countries in each cluster.



Delay (days)	Countries (ISO-2 Code)
Short (1)	PL, EG, RU, HR, MK, CI, LU, SK, IE, GB
Medium (9)	PH, CN, US, PK, AL, DK, FR, IL, CY, TR, BA
Long (16)	CH, AT, NO, RS

Figure 10: Three clusters are identified with Short (1 day), Medium (9 days) and Long (16 days) information delays. The table lists representative countries from each cluster.

5 Technical Implementation Details

Eva was deployed on Amazon Web Services (AWS) under an account maintained by the Ministry of Digital Governance, using servers and services that resided in Europe. The system consisted of four submodules:

1. The Application Communication Interface (API) used for communication with the Greek data provider (GR-DB)
2. The Eva Engine
3. The Database
4. The Continuous Deployment System

These submodules were contained in the same Virtual Private Cloud (VPC) and, to ensure privacy, no incoming communication to the open internet was allowed. Any incoming or outgoing data connections with GR-DB took place through a common VPC using the API submodule.

The *API submodule* operated on AWS Lambda, the serverless computing service of Amazon. The API had two serverless functions: one responsible for anonymized data ingestion, and the other for communication of results back to GR-DB. To provide additional security, token-based authentication was implemented and enforced, and all authentication attempts were logged.

The *Eva Engine* submodule operated on Amazon Elastic Compute Cloud (EC2) on an 8 V-CPU and 32 GiB Memory server. It contained the code for running the Eva algorithm, which outputs daily test allocations given recent testing results and the current passenger manifest. The results were stored for analysis purposes to the Database and were also transformed to various formats to be sent to the API.

The *Database* used the Amazon Serverless Aurora, and it was stored the incoming data, preprocessed results, and processed final results. It also kept track of which results were sent to GR-DB, effectively acting as a queuing system. The Database was accessed only from Eva's VPC.

A *Continuous Deployment system* was in place to ensure that updates to the code could be deployed easily and safely in the API or the Eva engine, while offering functionality for logging, notifications and rollback.

There were three identical versions of the Eva system all running in parallel: a production version that had the latest verified and approved code, a "staging" version that had code that had not yet been verified for validity, and a development version for active research and development. The daily cost of the Eva portion of the system ranged from $17 to $25 per day.