Interpretable Operations Research for High-Stakes Decisions: Designing the Greek COVID-19 Testing System

Hamsa Bastani
Wharton School, University of Pennsylvania, hamsab@wharton.upenn.edu

Kimon Drakopoulos, Vishal Gupta
USC Marshall School of Business, {drakopou, guptavis}@marshall.usc.edu

with

Jon Vlachogiannis*, Christos Hadjicristodoulou†, Gkikas Magiorkinis‡, Dimitrios Paraskevis‡, Pagona Lagiou‡, Sotirios Tsiodras‡

In the summer of 2020, in collaboration with the Greek government, we designed and deployed Eva – the first national scale, reinforcement learning system for targeted COVID-19 testing. In this paper, we detail the rationale for three major design/algorithmic elements: Eva's testing supply chain, estimating COVID-19 prevalence, and test allocation. Specifically, we describe the design of Eva's supply chain to collect and process thousands of biological samples per day with special emphasis on capacity procurement. Then, we propose a novel, empirical Bayes estimation strategy to estimate COVID-19 prevalence among different passenger types with limited data and showcase how these estimates were instrumental for a variety of downstream decision-making. Finally, we propose a novel, multi-armed bandit algorithm that dynamically allocates tests to arriving passengers in a non-stationary environment with delayed feedback and batched decisions. All of our design and algorithmic choices emphasize the need for transparent reasoning to enable human-in-the-loop analytics. Such transparency was crucial to building trust and buy-in among policymakers and public health experts in a period of global crisis.

Key words: targeted COVID-19 testing, human-in-the-loop analytics, supply chain design, empirical Bayes, LASSO, reinforcement learning

1. Introduction

In early 2020, countries restricted non-essential travel to mitigate the spread of the COVID-19 pandemic. These restrictions crippled many economies, with estimated losses of $1 trillion and 19 million jobs due to tourism in European countries alone (Wor20). As the first wave abated in April 2020, countries sought to reopen their borders, not only for tourists but also for goods and labor, all while continuing to safeguard public health.

Ideally, to prevent importing new cases, re-opening nations would test *every* arriving traveler and quarantine those who test positive and their contacts (AKOW20, FCZ20, RDJ20, KF20). Unfortunately, amidst the crisis, both testing resources and medical personnel were scarce, making this approach (even with group testing (Dor43)) untenable. Targeted allocation of scarce testing resources was necessary.

* AgentRisk † University of Thessaly ‡ National and Kapodistrian University of Athens

Against this backdrop, we partnered with the Greek Government to design and deploy Eva – the first national scale, reinforcement learning system that adapts to real-time information for targeted COVID-19 testing. Importantly, Eva uses testing results from recently screened passengers to learn and react to the evolving pandemic. Eva had two intertwined goals: i) optimize the allocation of Greece's scarce testing resources to identify as many infected, asymptomatic passengers at the border as possible, and ii) develop reliable estimates of COVID-19 prevalence across passengers from different origin countries to inform Greece's travel protocols and downstream decision-making.

Our partner paper (BDG$^+$21) describes the effectiveness of Eva and the resulting epidemiological insights; in contrast, the current paper delves into key operational and algorithmic design choices underlying Eva, with special emphasis on interpretability.

1.1. Overview of the EVA System

Eva as presented in this paper was deployed on August 6th, 2020 and remained in operation until October 30th, 2020 (at which point the tourist season ended and Greece returned to a lockdown). During the peak season (August, September), Eva processed travelers from approximately $41,830$ ($\pm 12,784$) households and performed $5,400$ (± 982) tests on a daily basis; accounting for no-shows, approximately 16.7% of daily arrivals were tested across 40 different ports of entry. The underlying supply chain included 120 teams of nurses and doctors, 200 firemen and policemen, 32 private and public testing labs, and logistics teams to transport biological samples from ports of entry to labs. To keep the current paper self-contained, we first briefly overview Eva's operations (see also Fig. 1):

1. *Passenger Locator Form (PLF)*: All travelers complete a PLF (one per household) at least 24 hours prior to arrival, with information on their origin country and region[1], age group, gender, point and date of entry, and intended destination within Greece.
2. *Estimating COVID-19 Prevalence among Traveler Types*: Using PLF data and recent testing results from previous travelers, we estimate the COVID-19 prevalence among different types of passengers. The procedure entails two steps: First, we group passengers into *types* (e.g., all passengers arriving from Madrid, Spain). Passengers in the same type are expected to have similar risk profiles. We then use empirical Bayes to estimate the prevalence for each type (see Sec. 4).

[1] "Region" is analogous to a state or province. Since different countries use different nomenclature, we use the generic "region."

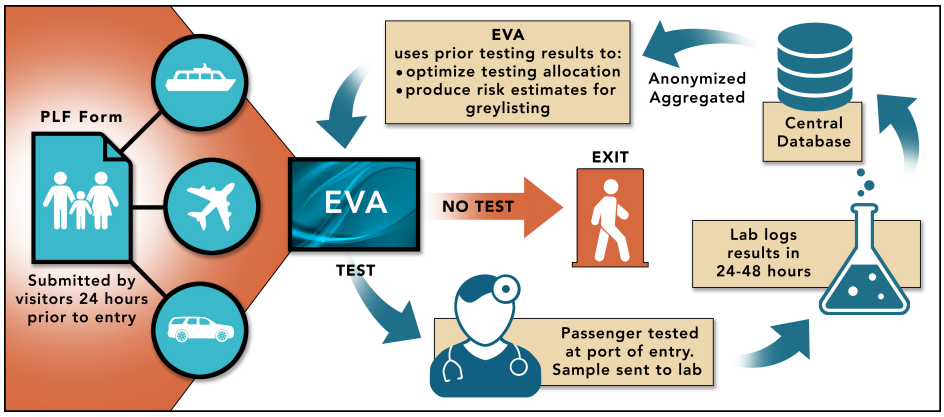

Figure 1 **Overview of Eva.** (Adapted from (BDG+21).) Eva, the first national-scale reinforcement learning system for targeted COVID-19 testing was deployed during the summer of 2020 across all 40 ports of entry to Greece. The closed-loop system reacts dynamically to the evolving pandemic to target tests, and was able to identify 1.85x as many asymptomatic, infected passengers at the border as a random policy.

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. This targeting must respect various operational constraints (e.g., delayed feedback from labs, resource constraints) that reflect Greece's testing supply chain (see Sec. 5).

   Allocations today determine the data available for estimating prevalence in the future. Hence, Eva strategically allocates some tests to traveler types for which it does not currently have precise prevalence estimates, to ensure better estimates in the future. This feedback step is crucial. If Eva simply greedily allocated tests to the highest risk passengers, after a few days, there would be no recent data on types with moderate prevalence. If COVID-19 prevalence then spiked in one of those moderate types, Eva would be blind to the increased risk to public health.

4. *Closing the Loop*: Anonymized results from the tests performed (Step 3) are added to the database in 24-48 hours, and used to update prevalence estimates (Step 2).

1.2. Impact: Identifying Asymptomatic, Infected Passengers

Before proceeding, we summarize some of the documented effectiveness of Eva from our partner paper (BDG+21). The left panel of Fig. 2 compares the number of asymptomatic, infected passengers caught by Eva during deployment to those a random, surveillance policy with the same testing budget would have caught. Random surveillance was the policy originally proposed by Greece (before our collaboration), because it requires no information

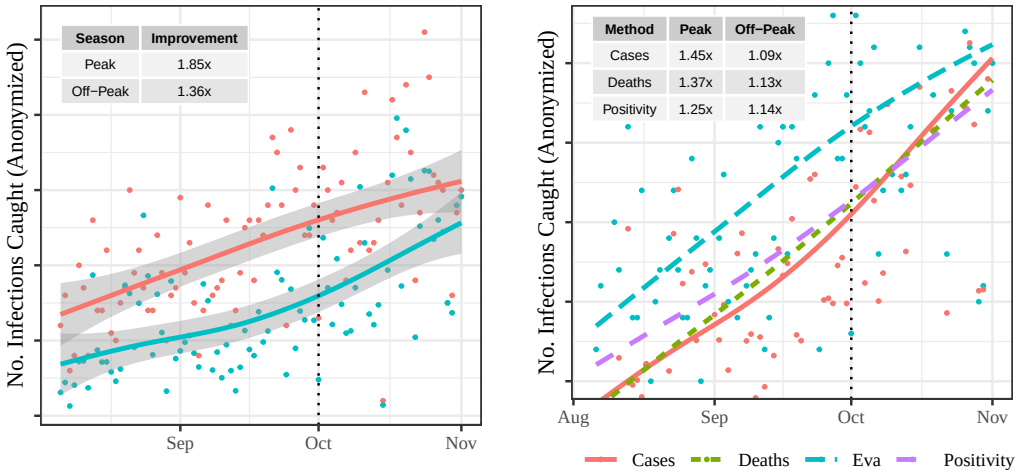

Figure 2 **Impact of Eva: Cases Identified Relative to Other Policies** (Adapted from (BDG$^+$21).) The left panel compares Eva to a random, surveillance policy; the right panel compares Eva to policies where the probability of testing a passenger is proportional to one of three epidemiological metrics for their origin country (cases per capita, deaths per capita, or positivity rate). The dotted vertical line separates the peak and off-peak travel seasons. Inset tables report the *relative* improvement of Eva over corresponding benchmark. The y-axis is anonymized for privacy reasons. Smoothed lines are computed by cubic-spline.

infrastructure to implement. Counterfactual analysis shows that, in the peak travel season (August, September), Eva identified approximately 1.85x as many infected, asymptomatic passengers as random surveillance testing (BDG$^+$21). In other words, Greece would have needed 85% more testing capacity to achieve the same effectiveness.

Similarly, the right panel compares performance to benchmark policies inspired by recent proposals by the EU (Gen20). These policies test travelers roughly proportionally to the "risk" of their origin country, where "risk" is defined using one of three widely-used epidemiological metrics: cases per capita, deaths per capita, positivity rates. Eva identified approximately 1.3x-1.5x as many infected, asymptomatic travelers in peak travel season (depending on the benchmark) (BDG$^+$21). Unlike random surveillance, these policies require similar information infrastructure investments as Eva, but achieve a fraction of the benefit. Thus, Eva's improvement relative to these policies stems solely from its ability to learn from recent testing results (reinforcement learning).

We refer the interested reader to (BDG$^+$21) for details of this analysis, and additional discussion of the epidemiological and public policy implications.

2. Contributions and Outline of Paper

Obviously, designing Eva involved a myriad of decisions. Many, if not all, of our design choices were shaped by the practical reality of earning trust and "buy-in" from a largely non-technical set of decision-makers in a period of crisis. This environment required that i) Eva's outputs, i.e., prevalence estimates and testing recommendations, follow from *transparent reasoning* and that ii) it was possible to easily incorporate domain expertise from epidemiologists and policy makers into any and all portions of the analytics pipeline, so-called *human-in-the-loop* analytics (BBK18, BA20, BBS21).

Indeed, some readers may recall that the early stages of the COVID-19 pandemic were characterized by immense uncertainty. Many basic epidemiological facts were still debated: Could the virus be transmitted by touching a contaminated surface? How long after infection was a patient contagious? What was the reliability of rapid testing vs. Polymerase Chain Reaction (PCR) testing? Similarly, there was little data (if any) to predict the public's response to the pandemic: How many would choose to travel if borders reopened? Would travelers' chosen destinations change? Finally, there was an incalculable number of "unknown unknowns" – unanswered questions we did not yet know to ask.

With such pervasive uncertainty, we argue that *transparent reasoning* and *human-in-the-loop analytics* are crucial to success. Hence, rather than laundry-list all the operations research models underlying Eva, we focus on three crucial elements of Eva's design and highlight how these two principles shaped our design decisions. Specifically, we discuss:

1. Designing Eva's Testing Supply Chain. Eva's ability to adapt to unexpected spikes in prevalence hinges on a robust supply chain that processes tests quickly. We formulate a stochastic, mixed-binary, linear optimization problem to model procuring testing capacity. Our formulation explicitly models the fact that tests will ultimately be allocated according to our targeting procedure, inducing decreasing marginal effectiveness of additional capacity at each port of entry. This is a distinctive and first-order feature of our application.

2. Estimating COVID-19 Prevalence Across Passenger Types. Accurate and intuitive prevalence estimates are key both for allocating tests to travelers as well as other downstream decisions by policy-makers. For example, Eva's prevalence estimates informed positioning of mobile-testing units within Greece (see Fig. 3), and also diplomatically sensitive, national-level travel protocols such as "grey-listing" (discussed in BDG$^+$21).

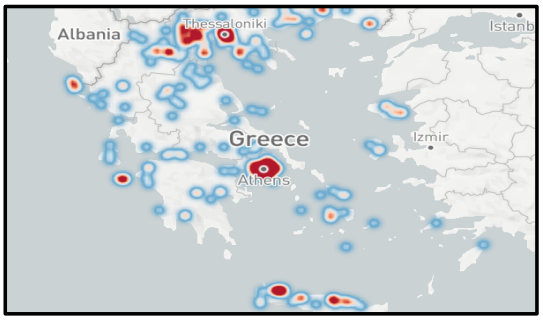

Figure 3 High Risk Locations within Greece. (Simulated data). By combining travelers' PLF data with Eva's prevalence estimates, we produced "risk" heat maps to guide within-country operations, including allocating mobile testing units to assess the spread to local populations in tourist-facing industries (e.g., hotels, restaurants, night-clubs), and adapt social distancing measures.

Unfortunately, effective estimation faces several statistical challenges: COVID-19 prevalence was relatively low in our population (on order of 0.2%). At the same time, arrival rates varied substantially across countries. Both features lead to *highly* imbalanced data (few positive cases and few arrivals from certain countries). Finally, prevalence is non-stationary, making old testing results uninformative.

We propose a novel estimation procedure for COVID-19 prevalence that combines LASSO regression and an empirical Bayes algorithm that seamlessly accounts for the aforementioned challenges around time-inhomogeneity, imbalanced data, and limited data for certain types. Perhaps surprisingly, our procedure yields estimates with simple, intuitive structure easily explained to non-technical audiences. We demonstrate empirically that our procedure is far more accurate than comparably simple maximum likelihood estimators, and substantially more stable than blackbox estimators like Gradient Boosted Machines. Our estimates achieve best-in-class performance when the number of origin countries is large, even if the number of travelers from each country is small, the so-called "small-data, large-scale" regime (GR21, GHR21, BSLZ19). In our non-stationary setting where old data is uninformative, this limit is more relevant than the traditional large-sample limit.

3. Adapting the Allocation of Tests to Real-Time Feedback. In Eva's closed-loop system, the passengers we choose to test today determine the available data for estimating COVID-19 prevalence in the future. Thus, when allocating tests, Eva must balance two competing goals: i) test "high-risk" passengers to prevent importing positive cases and ii) test all passengers to develop high-quality estimates across all types. These twin goals resemble the familiar "exploration-exploitation" tradeoff studied extensively in the multi-armed bandit literature (Tho33, LR85, Git79, Aue02). However, our setting exhibits a number of features that distinguish it from the classical setting: Labs require 24-48 hours to process results (delayed feedback), and lack of reliable internet connectivity at remote ports of entry

necessitate pre-computing all allocations at the beginning of each day (batched decision-making). Testing allocations must also respect supply chain constraints (constrained action sets). Existing multi-armed bandit algorithms either cannot handle all of these challenges, or perform excessive exploration in non-stationary environments.

Consequently, we propose a novel contextual multi-armed bandit algorithm for non-stationary environments under delayed feedback, batching, and constrained action sets. As discussed in Sec. 5, our algorithm uses a new technique we call "certainty-equivalent updates" to account for information we expect to receive in the future. We show that our algorithm leads to easily interpretable testing recommendations, unlike other popular contextual bandit algorithms, supporting our aim for transparent reasoning.

2.1. Connections to Existing Literature

Several papers have considered the allocation of scarce COVID-19 testing resources through simplified, theoretical models studying the structure of optimal testing policies (AFG$^+$21, CCI$^+$21, DR20, MZ21, CHZ20, KT20). By contrast, our work focuses on high-fidelity modeling in a reinforcement learning context, with a premium on transparency and interpretability. Moreover, to the best of our knowledge, we are the first to design and *deploy* such a policy at a large scale and provide empirical evidence of its advantages.

In some ways, our testing supply chain design problem resembles classical supply chain and network design problems like facility location and inventory fullfillment (AF19, AG15, DPW21, DGWZ16, MNSDG09). However, unlike (some of) these models, decisions in our setting are single-shot and non-adaptive. Moreover, demand will ultimately be served using our targeting procedure. This feature introduces an important non-linearity into the objective. Finally, our focus on "human-in-the-loop" analytics shifts focus away from finding an "optimal" supply chain and towards understanding how the design depends on hard-to-quantify costs to help decision-makers identify a good, robust design.

We utilize an empirical Bayes method for estimation. Empirical Bayes is common in settings where one must solve hundreds of separate statistical estimation problems simultaneously (Efr12, JRR20). In epidemiology, these methods have been used to estimate prevalence across many (potentially related) populations (GR91, DLH94). Recently, (IW19) proposed blending modern machine learning and empirical Bayes techniques. Our approach is closest to this last stream, since we use demographics and LASSO regression to first identify passenger types, and then empirical Bayes methods to form estimates.

Finally, our work contributes to the literature on multi-armed bandits. From an applied perspective, multi-armed bandits have been used in mobile health (TM17), clinical trial design (DAI$^+$18), online advertising (LCLS10) and recommender systems (ACJB18). To the best of our knowledge, however, ours is the first deployment of a multi-armed bandit algorithm at a national scale for a public-facing, high-stakes application.

From a theoretical perspective, there are various approaches to multi-armed bandit problems (Tho33, Git79, LR85, Aue02). Many authors have also sought to relax some stylized features of the classical setting including non-stationary rewards (BGZ14, LWAL18); high-dimensional contexts (BB20); batched decision-making (GHRZ19, PRC$^+$16); delayed-feedback (JGS13, VCL$^+$20); and budget constraints (AG13). Our algorithm addresses *all* of these elements. For interpretability, our approach builds on the optimistic gittins index approach (GF16), but with a new certainty-equivalent updating technique to account for batching/delayed feedback under non-stationarity.

3. Designing Eva's Testing Supply Chain

In the summer of 2020, neither the European Union nor the Greek COVID-19 Scientific committee had approved rapid tests due to concerns around accuracy. Consequently, Eva exclusively used Polymerase Chain Reaction (PCR) testing, which requires specialized equipment. Unfortunately, due to a rise in global demand, the time to procure and set-up new equipment was estimated to be 3-6 months, making capacity expansion impossible.

Instead, Greece contracted testing capacity with 32 private and public labs to process biological samples. Labs were paid per test processed, with no significant differences in pricing. Based on their equipment and personnel, each lab had its own maximum daily processing capacity and required biological samples collected with either a dry-swab or a wet-swab. Different equipment is required for each swab type, and *not* all labs can analyze both swab types. Procurement costs for swabs are negligible relative to processing costs.

Tests were performed at each port of entry by dedicated medical personnel. Ports of entry that used both swab types faced more complex operations because the types are handled differently (DGT$^+$12). Quantifying the cost of this complexity is difficult. Hence, policy-makers preferred to use one swab type at each port of entry "as much as possible."

Transporting biological samples safely to labs also requires special equipment. Since Greece has many points of entry on remote islands and samples must be transported twice

daily to maintain fast processing times, a given port of entry can only be served by a subset of nearby labs. Because of the (relatively) short distances traveled, the dominant transportation costs were fixed "per trip" costs that were similar across all entry/lab pairs.

Finally, demand far exceeds the supply in this system; we could only test about 17% of arrivals and our allocation procedure fully utilizes capacity by design. Issues around flexibly allocating spare capacity to handle excess demand (as in the celebrated long-chain, JG95) were largely second-order. Thus, to the extent possible, policy-makers preferred designs where each port of entry was served by a single lab to minimize the chance of a "cascading" back-log across labs or an error in shipment.

In summary, inventory and procurement costs are negligible, shipment costs are largely homogeneous "per-trip", and per-unit processing costs are homogeneous. Moreover, adjusting the supply chain mid-summer was deemed too expensive and risky, and so design decisions (particularly with outside private contractors) were irrevocable. Therefore, we model the resulting supply chain design problem as a single-stage stochastic, mixed-binary linear optimization problem (see Appendix A). The primary decision variables encode how many tests each lab will process from each port of entry, subject to i) only processing tests from sufficiently nearby ports of entry, ii) not exceeding lab processing capacities and iii) ensuring ports of entry use swab types compatible with their respective labs. Notice that given a feasible solution to this problem, one can immediately calculate the available testing capacity at each port of entry (an input to the allocation procedure) and the corresponding number of required medical personnel to support that capacity.

Specifying an Objective. Intuitively, a good supply chain design will maximize the expected number of infections caught by our targeting procedure throughout the summer, while penalizing excessive shipment costs and the number of ports using both swab types. Formalizing this intuition into an objective faces two challenges: i) For a given testing capacity at a port of entry, how do we quantify the number of infections that will be caught by our targeting policy? ii) What is the relative benefit of finding more infections compared to reducing shipment costs or reducing the number of ports using both swab types?

For the first challenge, traditional supply chain models assume a fixed benefit per-unit of demand served, suggesting a constant marginal benefit for each additional test allocated to a port. Some reflection suggests, however, the marginal benefit of serving more demand at a port of entry is decreasing under targeted testing; the first few tests will be given

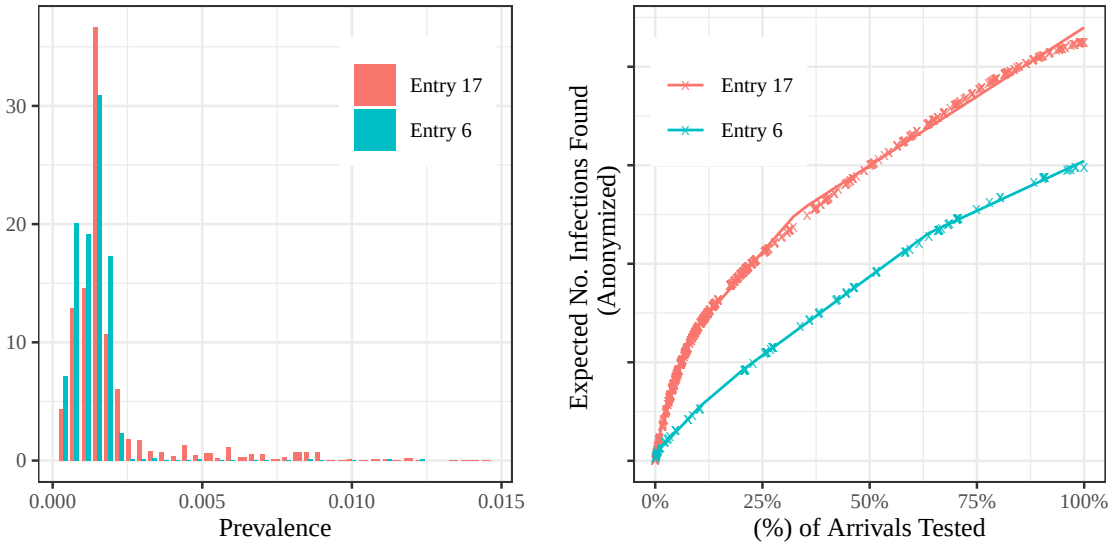

Figure 4 **Contrasting Risk at Ports of Entry.** The left panel contrasts the distribution of COVID-19 risk among arriving passengers at two different ports of entry. The right panel shows the expected number of infections caught when varying the percentage of arrivals tested at each port (marked with an "x".) Arrival heterogeneity induces very different functions. Solid lines show our piecewise linear approximation.

to the highest-risk passengers, and additional tests necessarily go to lower-risk passengers. Moreover, because the mix of types of passengers at each port varies substantially, the prevalence distribution among arriving passengers also varies substantially, causing the benefit of serving additional demand to also vary. Fig. 4 depicts a concrete example contrasting the prevalence distribution at two ports of entry.

We show in Appendix A that the expected number of infections caught at a port under perfect targeted testing is a concave function of the fraction of arrivals tested (see right panel of Fig. 4). Although one can maximize this concave function using lazy cuts and outer-linearization (see Appendix A), we prefer a simpler approach and approximate the relevant function by a piecewise-linear concave function with a small number of pieces. The right panel of Fig. 4 shows that we already obtain an accurate fit with 3 pieces.

Our approach to the second challenge – quantifying the benefit of identifying additional infections relative to shipping costs or the increased complexity of stocking both swab types – was again inspired by human-in-the-loop thinking. Policy-makers were reluctant or unable to quantify these costs which fundamentally compare public health to money spent. Rather than engage in a difficult preference elicitation exercise, we built a decision-support tool that allowed policy-makers to compare the expected number of infections caught by a supply chain design and the actual structure of the design as we varied these parameters.

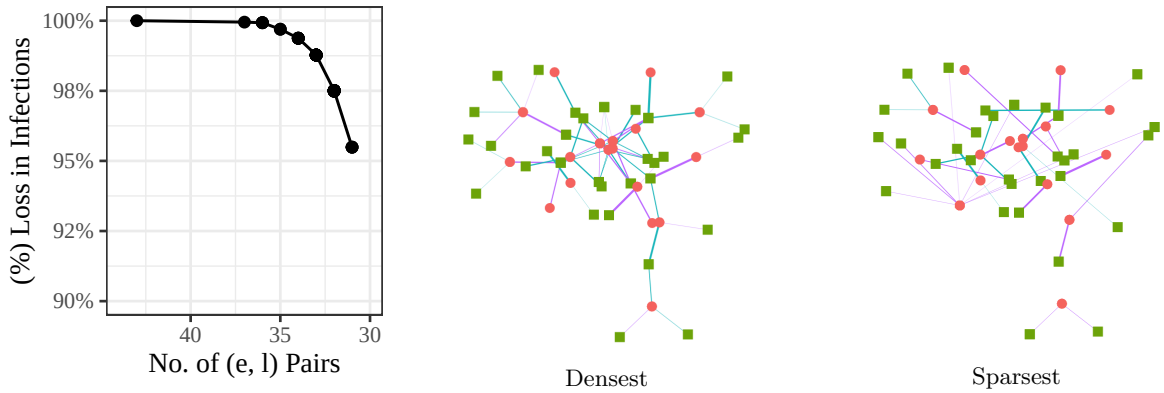

Figure 5 **Infections vs. Shipment Cost Tradeoff.** First Panel: As the number of entry/lab pairs used grows, shipping costs increase proportionally, but the expected number of infections identified only increases marginally. Y-axis is relative to the number of infections caught by the densest configuration. Second Panel: Structure of sparsest chain (31 routes), chosen to ensure that no ports of entry are left unstaffed. Third Panel: Structure of densest chain (43 routes). Red circles represent labs, green squares represent ports of entry, and blue (resp. red) edges represent the number of wet (resp. dry) swabs. Edge weights are proportional to the number of tests processed. Results depict a subset of 18 labs and 31 ports of entry.

For example, Fig. 5 shows the tradeoff between the number of infections caught and the number of entry/lab pairs in the chain (which is proportional to shipping costs). It also shows two designs, corresponding to the sparsest and densest configurations.

As we use more routes (and shipping costs increase), we observe only a marginal increase in the expected number of infections identified. The structure of the resulting network is somewhat different. By examining this tradeoff and the underlying network, decision-makers were able to incorporate their own expertise and "on-the-ground" experience to select a suitable design, without needing to quantify their relative preference between infections caught and Euros saved.

A similar analysis shows that using only one swab type per port of entry results in no loss in number of infections identified.

4. Estimating COVID-19 Prevalence

We next describe our procedure for estimating COVID-19 prevalence with special emphasis on interpretability. To be concrete, in what follows we focus on one, specific property that we argue interpretable estimators should satisfy in our setting:

Definition An estimator enjoys *perturbation monotonicity* if, after adding exactly 1 positive case (and 0 negative cases) to the data, no type's prevalence estimate decreases.

We argue that decision-makers intuitively expect perturbation monotonicity. Estimates are updated daily, and epidemiologists on the Greek COVID-19 taskforce monitor trends. If, after observing one positive from a woman 30-40 years old from Country A and *no other data*, the prevalence estimates for men 20-30 years old from Country B (a seemingly unrelated type) suddenly *decreased*, questions are naturally raised. Said differently, estimators that do not satisfy perturbation monotonicity are not transparent.

We next show that some state-of-the-art machine learning methods do not guarantee perturbation monotonicity.

4.1. Drawbacks of Blackbox Models

Gradient Boosted Machines (GBM) regression is one of the most robust and accurate blackbox machine learning methods for structured data (Had16). Unfortunately, GBM does not satisfy perturbation monotonicity.

To illustrate, we trained a GBM to predict prevalences using passenger PLF data (age group, gender, country of origin[2]) and testing data from Eva from the two weeks preceding Sept. 1, 2020.[3] There are 67,154 data points in this time frame. We then perturbed these data by adding a *single* positive case for a synthetic passenger from country A (anonymized) with other features drawn randomly. We retrain the GBM on the augmented dataset and compare estimates (see Fig. 6).

Prevalence estimates (averaged over all passengers from a country) move in both directions, sometimes dramatically. Across all countries, the estimated changes ranged from 81% reduction (Country B) to a 504% increase (Country G). Practically, such dramatic changes from *one* additional case are untenable. For example, before the additional case, Country B exhibited a prevalence above 3%, high enough to merit discussions about changing national-level, travel protocols with that country. After the addition of a positive case from country A, a seemingly unrelated type, Country B's estimate dropped substantially. Given the range of downstream activities informed by these estimates, such swings could be quite dangerous.

Although our discussion above focuses on GBM, similar behavior is common in other blackbox models. Their lack of transparent reasoning makes diagnosing the cause of, and explaining the reason for, problematic behaviors challenging.

[2] We exclude region of origin in this exercise as there were 17,000 distinct regions in the data, requiring 17,000 features in a standard one-hot encoding. The R implementation of GBM supports categorical variables with at most 1,024 unique values. Intuitively, including more features only *increases* instability. [3] We only use recent data for estimation because prevalence is non-stationary (old data is uninformative). Conversations with epidemiologists suggested that a 2-week window was appropriate given known evidence at the time.

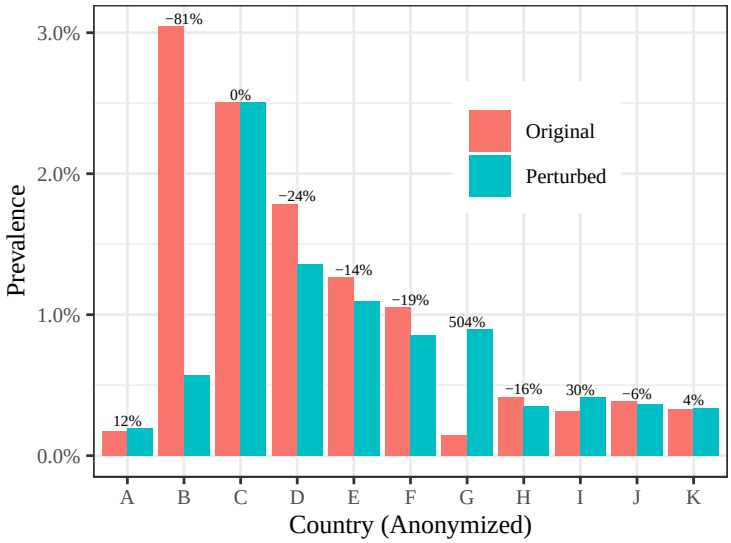

Figure 6 Instability of GBM.
Average estimated prevalence for passengers from Country A and the top 10 highest-risk countries as of Sept. 1, 2020 using GBM (red bars). Green bars show estimates after observing one additional positive case from country A. Labeled percentages show the change in estimates after the perturbation. GBM does not perturbation monotonicity property.

4.2. Our Empirical Bayes Approach

Unlike GBM, many simple estimators do enjoy perturbation monotonicity. Consider the maximum likelihood estimator $\hat{r}^{MLE}(t,c)$ which estimates the probability that a passenger from country c is infected by the proportion of positive tests among passengers from c in the last two weeks. By construction, if we add one additional case from country A, the estimate for country A increases and all other estimates remain unchanged.

Unfortunately, the maximum likelihood estimator (MLE) suffers from imbalanced data, particularly for rare countries. As an illustration, we compare $\hat{r}^{MLE}(t,c)$ to a (foolish) baseline estimator $\hat{r}^{Base}$, which estimates 0 prevalence for all passengers. A leave-one-out type approximation to the mean-squared error (MSE) using the same data from Eva as above suggests the excess MSE of $\hat{r}^{MLE}(t,c)$ – i.e., the difference between the MSE of $\hat{r}^{MLE}(t,c)$ and $\hat{r}^{Base}$ – is .000334. The MLE is worse than the baseline! In fact, the typical error in $\hat{r}^{MLE}$ is *at least* the square-root of this excess MSE, i.e., $\sqrt{.000334} = 0.018$. Since the typical prevalence in our data set is 0.002, this suggests that the noise in the MLE completely washes out any potential signal.

Consequently, we adopt an empirical Bayes perspective which is far more interpretable than blackbox methods, but still performs well with imbalanced data. We group passengers into distinct "types," indexed by k, and estimate prevalence separately for each type. On first reading, the reader can interpret a passenger's type as their country of origin (see Sec. 4.3 below for details on type definition.)

A traditional Bayesian model might assert that the prevalence for type k is drawn from a common prior distribution for all k, and, conditional on that draw, recent passengers are positive with probability given by the draw. Importantly, in this model, the prior is chosen based on domain knowledge alone, without leveraging the observed data. One then computes the posterior distribution using the data, and estimates prevalence with the posterior mean. If the prior is a Beta(α, β) distribution, the posterior mean is

$$\hat{r}^{Bayes}(t,k) = \frac{S_0}{S_0 + n(t,k)} r^{Prior} + \frac{n(t,k)}{S_0 + n(t,k)} \hat{r}^{MLE}(t,k), \tag{1}$$

where $r^{Prior} = \frac{\alpha}{\alpha+\beta}$ is the prior mean, $\hat{r}^{MLE}(t,k)$ is the observed proportion of positives among type k passengers in the past two weeks, $n(t,k)$ is the number of type k passengers tested in the past two weeks, and $S_0 = \alpha + \beta$ is the strength of the prior.

Our empirical Bayes procedure is very similar to the classical procedure conceptually. However, instead of specifying α, β (equivalently r^{prior}, S_0) from domain knowledge, we fit them from the data. We adopt a moment-matching approach common in the empirical Bayes literature; we estimate $\hat{r}^{prior}$ and $\hat{S}_0$ in order to match the first two moments of the observed testing data. We then plug these estimated values into Eq. (1) to obtain our estimate $\hat{r}^{EB}(t,k)$ prevalence (see Appendix B for details).

The resulting estimator has an intuitive form. Eq. (1) is a convex combination between the MLE and the prior mean. For types with large $n(t,k)$ (lots of data), the convex combination is close to the MLE, which we would intuitively expect. For types with small $n(t,k)$ (limited data), the estimate shrinks towards the prior mean r^{prior}, ameliorating the aforementioned variability for rare types. Hence, types with limited data "borrow information" from types with more data to reduce variability. Moreover, our moment-matching construction chooses $\hat{r}^{prior}$ to be the average observed prevalence across all types, i.e., estimates for rare types are shrunk towards the "typical" prevalence.

Fig. 7 illustrates this idea using the same data from Eva used above. Here, we fit the prior mean via our moment-matching method, and plot the excess MSE over the baseline estimator r^{Base} averaged over all types k at time t as we vary S_0. From Eq. (1), $S_0 = 0$ corresponds to the MLE $r^{MLE}(t,k)$, but its excess MSE is too large for the plot. The red point corresponds to our estimator $\hat{r}^{EB}(t,k)$, which substantially improves upon $\hat{r}^{MLE}(t,k)$.

Our estimator does not directly minimize the estimate of MSE. We prefer our procedure to directly minimization because, as seen in Appendix B, our estimate admits a simple

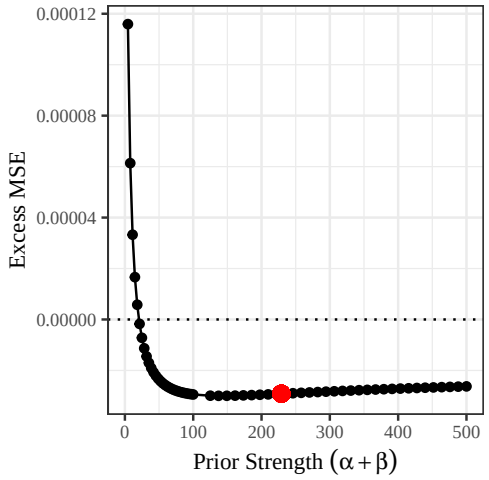

Figure 7 Excess MSE over baseline estimator.

The x-axis depicts the prior strength $S_0 = \alpha + \beta$. The MLE estimate $\hat{r}^{MLE}$ (clipped from plot for readability) corresponds to a prior strength of zero and has far worse MSE than the baseline of always estimating zero. Our empirical Bayes estimator (red dot) substantively improves performance and approximately minimizes MSE by trading off bias and variance.

closed-form formula that is very robust to data perturbations when the number of types is large. Indeed, we prove that as the number of types grows to infinity, our estimator enjoys perturbation monotonicity, and converges to the Bayes optimal estimator even if the number of tests $n(t,k)$ performed for each type is small. This strong performance guarantee mirrors similar guarantees in the so-called small-data, large-scale regime (GR21).

4.3. Identifying Interpretable Types

We next describe our method for choosing a small set of interpretable types. PLF data specify a passenger's age group (ten buckets), gender (3 buckets), country of origin, and region of origin. In a typical two-week period, Eva saw arrivals from approximately 170 distinct countries and 17,000 regions. Thus, with the most granular definition, we would need $10 \times 3 \times 17{,}000 > 500{,}000$ types.

Statistically, estimation in such high dimensions is challenging. Practically, such a large number of types gives little insight to decision-makers. Consequently, inspired by the screening literature (FL08), we limit the number of types under consideration.

Our procedure has three steps: First, we fit our empirical Bayes estimates assuming types are defined by country of origin. This approach accords with epidemiological intuition because geography is highly predictive of prevalence. Second, we fit a LASSO logistic regression to predict a passenger's test results based on the fitted empirical Bayes prevalence for their country of origin, and dummy variables encoding their gender, age group and region of origin. The resulting non-zero components identified by LASSO signal buckets (e.g., regions) where testing results differed significantly from what one would expect based on origin country alone. Third, we re-define types more granularly, creating new

types for each of the non-zero LASSO coefficients, and refit our empirical Bayes estimates with the new set of types. For example, if the dummy variable for Madrid, Spain was non-zero in the LASSO fit, we removed the type for "Spain," and replaced it with "Spain, Madrid" and "Spain, not Madrid." After all such replacements, we re-run our empirical Bayes procedure with the larger set of types (see also (BDG$^+$21) for additional details).

Our procedure ensures that significant signals beyond country of origin are isolated and exploited. Over the summer of 2020, this procedure never identified additional types based on age or gender, and only identified a handful of types based on specific regions. In this sense, the number of types constructed was minimal. Having a smaller number of types defined solely in terms of origin country and region improves interpretability, allowing public health officials to sanity-check estimates against other sources of data.

4.4. Communicating with Policy-Makers

Fig. 8 shows the resulting empirical Bayes estimates for prevalence with confidence intervals for a typical day from the summer of 2020. This plot appeared on the dashboards of Greek officials to summarize the status of the pandemic and inform downstream decisions.

A feature of our empirical Bayes procedure is that we obtain posterior distributions and, hence, can easily illustrate (approximate) confidence intervals. Such confidence intervals helped us communicate uncertainty to non-technical collaborators, and, in particular, the need for exploration tests, i.e., tests allocated to types with moderate prevalence types but high uncertainty. Indeed, some collaborators initially viewed such tests as wasteful. Visualizations like Fig. 8 illustrated that large confidence intervals might be hiding risky types and required additional tests to resolve uncertainty. We discuss the test allocation procedure and how we ensured its transparency in the next section.

5. Allocating Tests Transparently

Finally, we describe our strategy for allocating tests based on our prevalence estimates.

5.1. Certainty-Equivalent Updates

Two key challenges in our setting are batching (all testing decisions for the day must be pre-computed at the start of the day) and delayed feedback (it takes 24-48 hours for labs to process tests). Consequently, at any given point of time, there are a large number of *pipeline tests* – tests that have been allocated but whose results have not yet been observed.

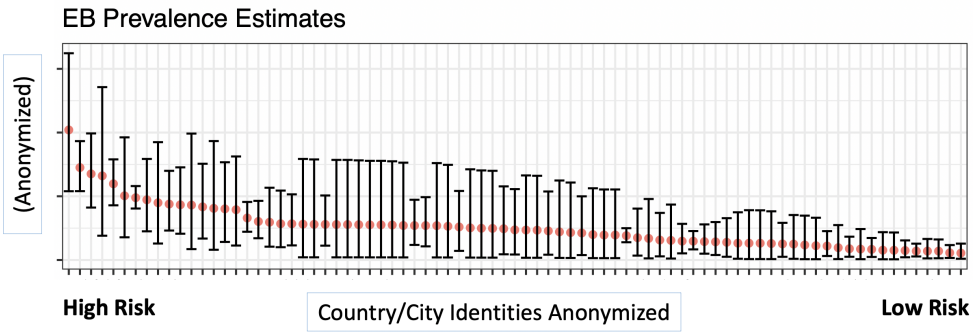

Figure 8 **Prevalence Estimates.** (Adapted from (BDG+21)) Actual estimated prevalence and 95% confidence intervals for each type on a given day. The x-axis indexes different types (anonymized) ordered from highest prevalence (left) to lowest prevalence (right). This plot was featured on policymakers' dashboards.

Standard approaches simply ignore pipeline tests and update estimates only when the results become available (see, e.g., JGS13, VCL+20). However, this strategy is problematic since it leads to significant *over-exploration*, which can hurt performance and erode trust. In particular, suppose type k passengers have an estimated prevalence that is low but uncertain. In the classical bandit setting, we would allocate a *few* exploration tests to type k passengers, use the resulting feedback to reduce the uncertainty of type k estimates, and then exploit by allocating remaining tests to passenger types with high estimated prevalence. As noted in the previous section, we can effectively communicate this need for limited exploration of uncertain types using Fig. 8. However, in the presence of delayed feedback and batched decision-making, we would *keep* allocating tests to type k passengers until we observe feedback that reduces the uncertainty, which might take two days and potentially thousands of tests. Consequently, we would allocate far more exploration tests to type k passengers than necessary to resolve uncertainty, thereby missing out on the opportunity to exploit (catch likely infected travelers at the border).[4] Naturally, this hurts performance and erodes trust in the allocation mechanism.

We introduce a novel algorithmic technique called *certainty-equivalent (CE) updates* to effectively balance exploration and exploitation despite the large number of pipeline tests. Although we do not observe immediate feedback when allocating a test, we can still estimate the likely reduction in the variance of our posterior distributions. If we ignore

[4] The batched bandit literature (PRC+16, GHRZ19) 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 prevalence.

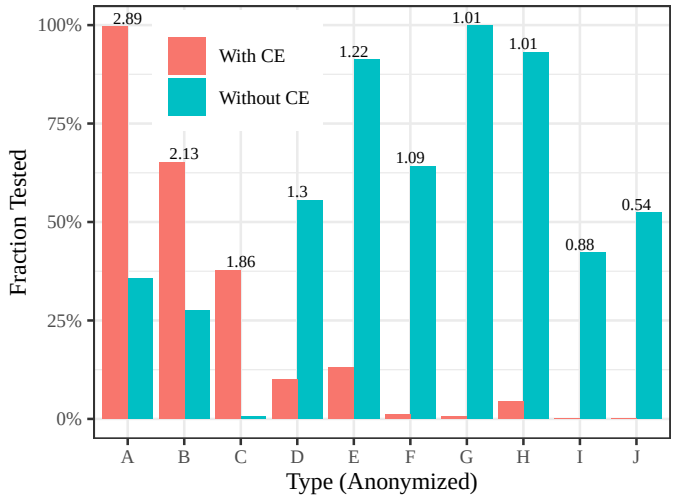

Figure 9 Over-exploration without CE updates. Fraction of arrivals tested for 10 countries where the two policies differ on Sept. 1, 2020 using our bandit algorithm with and without CE updates (red and green bars, respectively). Labeled numbers show estimated prevalence per 1000. Without CE updates, the bandit over-explores types with low prevalence; with CE updates, it allocates a few tests to all types (exploration) but primarily tests types with high prevalence (exploitation).

the uncertainty (motivating our nomenclature certainty-equivalent) around our estimate $\hat{r}^{EB}(t,k)$, then, after allocating an additional test to a type k passenger, we expect to observe a positive result with probability $\hat{r}^{EB}(t,k)$ and a negative result with probability 1-$\hat{r}^{EB}(t,k)$ within 48 hours. Accordingly, after allocating a test, we update the parameters of our posterior distribution with this expected result (see Appendix C). The update does not change our estimate of the mean prevalence, but reduces the variance of our posterior distributions. In this sense, it quantifies the expected additional information we will receive when test results are returned. A related heuristic was proposed by (BM07) for upper confidence bound (UCB) algorithms. Importantly, CE updates cause our optimistic gittins indices to also update as we allocate tests within a batch (day). Thus, they ensure that our algorithm allocates a *minimal* 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).

To illustrate, we computed test allocations on Sept. 1, 2020 using PLF and testing data from Eva. Fig. 9 shows the resulting allocations for 10 (anonymized) countries with and without CE updates. Without CE updates, the algorithm allocates an excessive number of tests to passengers from countries with low prevalence; using CE updates ensures that the majority of tests are allocated to countries with high prevalence, and only a fraction of exploratory tests are allocated to countries with low prevalence. A particularly striking example is contrasting types A and J (anonymized). The prevalence of type A is 2.89 cases per thousand, whereas the prevalence of type J is more than five times smaller. However, the strategy without CE updates tests over 50% of passengers of type J but only about

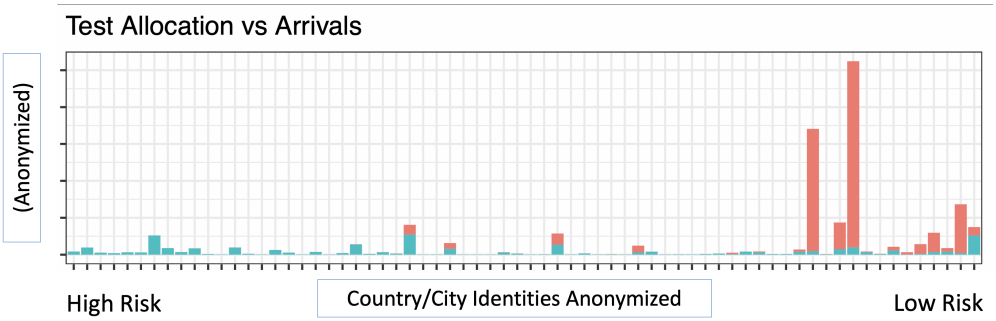

Figure 10 **Bandit Test Allocations.** (Adapted from (BDG+21).) Actual number of test allocations (teal) and untested scheduled arrivals (pink) for each type on a given day. The x-axis indexes different types ordered from highest prevalence (left) to lowest prevalence (right). This plot was featured on policymakers' dashboards.

30% of passengers of type A. In contrast, the strategy with CE updates tests a very small fraction of passengers of type J but nearly 100% of passengers of type A. These issues not only affect the performance of our algorithm, but can also be detrimental to building trust with decision-makers — i.e., excessively testing seemingly safe types raises questions.

Fig. 10 shows our test allocations for the same data as Fig. 8. This plot was also featured on the dashboards of policy-makers in the Greek government. Our algorithm allocates tests to essentially all high-risk arrivals (exploitation) but additionally assigns a small number of tests to all types (exploration).

5.2. Optimistic Gittins Indices: An Interpretable Risk Metric

We incorporate the above CE procedure into the optimistic gittins index algorithm of (GF16) to determine testing allocations. We prefer gittins indices to other algorithms, such as Thompson Sampling, partially because gittins indices are interpretable, deterministic risk estimates that capture both the immediate utility of testing a passenger (i.e., their estimated prevalence) as well as the future utility from reducing uncertainty in their estimate. In our implementation, we estimate our indices with a short (1-step) look-ahead window, i.e., we only consider the future utility of exploration for one additional day. This choice was motivated by our highly non-stationary environment since data collected for exploration are only useful for making predictions for a very limited future horizon.

5.3. Accounting for Supply Chain Constraints

Our test allocations are pre-computed at the start of the day. We first estimate the posterior distributions of each type using our empirical Bayes strategy, and perform CE updates

to account for the expected information of current pipeline tests. We then sequentially allocate a test to the type with the highest optimistic gittins index for which there are still available untested passengers, perform a CE update to the posterior for that type, and repeat until we deplete our testing budget or run out of passengers.

Recall that our allocation mechanism must satisfy constraints on port-specific testing budgets and arrivals. In particular, once our bandit algorithm identifies a passenger type to test, we must decide *where* to allocate the test; depending on the choice, it will consume testing budget at a particular port of entry. Some ports of entry have very limited testing budgets, and passengers of some types only travel to certain ports of entry. Intuitively, one should not allocate tests to common passenger types at a port of entry predominantly used by rare types. We employ a greedy heuristic that strategically "saves tests" at ports with few remaining tests for arrivals of rare types (see Appendix C).

6. Conclusions

We presented the rationale behind key algorithmic design choices for Eva, Greece's COVID-19 targeted testing system. In our view, similar principles should guide the deployment of any high-stakes, public-facing decision-making system. Such systems necessarily involve a wide network of stakeholders (e.g., policymakers, public health experts). Thus, it is critical that the system produce interpretable outputs and recommendations supported by transparent reasoning in order to promote human-in-the loop decision-making with domain experts.

Acknowledgments

The authors 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 thank Osbert Bastani for helpful discussions. V.G. was partially supported by the National Science Foundation through NSF Grant CMMI-1661732.

References

[ACJB18] Fernando Amat, Ashok Chandrashekar, Tony Jebara, and Justin Basilico. Artwork personalization at Netflix. In *Proceedings of the 12th ACM Conference on Recommender Systems*, pages 487–488, 2018.

[AF19] Jason Acimovic and Vivek F Farias. The fulfillment-optimization problem. In *Operations Research & Management Science in the age of analytics*, pages 218–237. INFORMS, 2019.

[AFG+21] Daron Acemoglu, Alireza Fallah, A. Giometto, D. Huttenlocher, A. Ozdaglar, F. Parise, and S. Pattathil. Optimal adaptive testing for epidemic control: combining molecular and serology tests. *ArXiv*, abs/2101.00773, 2021.

[AG13] Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *International Conference on Machine Learning*, pages 127–135. PMLR, 2013.

[AG15] Jason Acimovic and Stephen C Graves. Making better fulfillment decisions on the fly in an online retail environment. *Manufacturing & Service Operations Management*, 17(1):34–51, 2015.

[AKOW20] Ned Augenblick, Jonathan T Kolstad, Ziad Obermeyer, and Ao Wang. Group testing in a pandemic: The role of frequent testing, correlated risk, and machine learning. Technical report, National Bureau of Economi Research, 2020.

[Aue02] Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422, 2002.

[BA20] Margrét Vilborg Bjarnadóttir and David Anderson. Machine learning in healthcare: Fairness, issues, and challenges. In *Pushing the Boundaries: Frontiers in Impactful OR/OM Research*, pages 64–83. INFORMS, 2020.

[BB20] Hamsa Bastani and Mohsen Bayati. Online decision making with high-dimensional covariates. *Operations Research*, 68(1):276–294, 2020.

[BBK18] Hamsa Bastani, Osbert Bastani, and Carolyn Kim. Interpreting predictive models for human-in-the-loop analytics. *arXiv preprint arXiv:1705.08504*, pages 1–45, 2018.

[BBS21] Hamsa Bastani, Osbert Bastani, and Wichinpong Park Sinchaisri. Improving human decision-making with machine learning. *arXiv preprint arXiv:2108.08454*, 2021.

[BDG+21] Hamsa Bastani, Kimon Drakopoulos, Vishal Gupta, Jon Vlachogiannis, Christos Hadjicristodoulou, Pagona Lagiou, Gkikas Magiorkinis, Dimitrios Paraskevis, and Sotirios Tsiodras. Efficient and targeted covid-19 border testing via reinforcement learning. *SSRN*, 2021.

[BGZ14] Omar Besbes, Yonatan Gur, and Assaf Zeevi. Stochastic multi-armed-bandit problem with non-stationary rewards. *Advances in Neural Information Processing Systems*, 27:199–207, 2014.

[BM07] Dimitris Bertsimas and Adam J Mersereau. A learning approach for interactive marketing to a customer segment. *Operations Research*, 55(6):1120–1135, 2007.

[BSLZ19] Hamsa Bastani, David Simchi-Levi, and Ruihao Zhu. Meta dynamic pricing: Transfer learning across experiments. *Available at SSRN 3334629*, 2019.

[CCI+21] Sergio Camelo, Dragos F. Ciocan, Dan A. Iancu, Xavier S. Warnes, and Spyros I. Zoumpoulis. Quantifying the benefits of targeting for pandemic response. *medRxiv*, 2021.

[CHZ20] Ningyuan Chen, Ming Hu, and Chaoyu Zhang. Capacitated sir model with an application to covid-19. *Available at SSRN 3692751*, 2020.

[DAI+18] Audrey Durand, Charis Achilleos, Demetris Iacovides, Katerina Strati, Georgios D Mitsis, and Joelle Pineau. Contextual bandits for adapting treatment in a mouse model of de novo carcinogenesis. In *Machine Learning for Healthcare Conference*, pages 67–82. PMLR, 2018.

[DGT+12] Julian Druce, Katherine Garcia, Thomas Tran, Georgina Papadakis, and Chris Birch. Evaluation of swabs, transport media, and specimen transport conditions for optimal detection of viruses by pcr. *Journal of clinical microbiology*, 50(3):1064–1065, 2012.

[DGWZ16] Antoine Désir, Vineet Goyal, Yehua Wei, and Jiawei Zhang. Sparse process flexibility designs: Is the long chain really optimal? *Operations Research*, 64(2):416–431, 2016.

[DLH94] Owen J Devine, Thomas A Louis, and M Elizabeth Halloran. Empirical Bayes methods for stabilizing incidence rates before mapping. *Epidemiology*, pages 622–630, 1994.

[Dor43] Robert Dorfman. The detection of defective members of large populations. *The Annals of Mathematical Statistics*, 14(4):436–440, 1943.

[DPW21] Levi DeValve, Saša Pekeč, and Yehua Wei. Approximate submodularity in network design problems. *Available at SSRN 3844987*, 2021.

[DR20] Kimon Drakopoulos and Ramandeep S Randhawa. Why perfect tests may not be worth waiting for: Information as a commodity. *Available at SSRN 3565245*, 2020.

[Efr12] Bradley Efron. *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, volume 1. Cambridge University Press, 2012.

[FCZ20] P Frazier, M Cashore, and Y Zhang. Feasibility of COVID-19 screening for the US population with group testing. Technical report, Technical report, Cornell University, 2020.

[FL08] Jianqing Fan and Jinchi Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911, 2008.

[Gen20] General Secretariat of the Council. Draft council recommendation on a coordinated approach to the restriction of free movement in response to the COVID-19 pandemic, Brussels. Technical report, European Commission, Directorate-General for Justice and Consumers, 2020.

[GF16] Eli Gutin and Vivek Farias. Optimistic gittins indices. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

[GHR21] Vishal Gupta, Michael Huang, and Paat Rusmevichientong. Debiasing in-sample policy performance for small-data, large-scale optimization. *Large-Scale Optimization (June 2, 2021)*, 2021.

[GHRZ19] Zijun Gao, Yanjun Han, Zhimei Ren, and Zhengqing Zhou. Batched multi-armed bandits problem. *arXiv preprint arXiv:1904.01763*, 2019.

[Git79] John C Gittins. Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society: Series B (Methodological)*, 41(2):148–164, 1979.

[GR91] Sander Greenland and James M Robins. Empirical Bayes adjustments for multiple comparisons are sometimes useful. *Epidemiology*, pages 244–251, 1991.

[GR21] Vishal Gupta and Paat Rusmevichientong. Small-data, large-scale linear optimization with uncertain objectives. *Management Science*, 67(1):220–241, 2021.

[Had16] Sally Hadidi. Anthony Goldbloom gives you the secret to winning Kaggle competitions, Jan 2016.

[IW19] Nikolaos Ignatiadis and Stefan Wager. Covariate-powered empirical bayes estimation. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[JG95] William C Jordan and Stephen C Graves. Principles on the benefits of manufacturing process flexibility. *Management science*, 41(4):577–594, 1995.

[JGS13] Pooria Joulani, Andras Gyorgy, and Csaba Szepesvári. Online learning under delayed feedback. In *International Conference on Machine Learning*, pages 1453–1461. PMLR, 2013.

[JRR20] Gareth M James, Peter Radchenko, and Bradley Rava. Irrational exuberance: Correcting bias in probability estimates. *Journal of the American Statistical Association*, pages 1–14, 2020.

[KF20] Edward H Kaplan and Howard P Forman. Logistics of aggressive community screening for coronavirus 2019. In *JAMA Health Forum*, number 5, pages e200565–e200565. American Medical Association, 2020.

[KT20] Maximilian Kasy and Alexander Teytelboym. Adaptive targeted infectious disease testing. *Oxford Review of Economic Policy*, 36(Supplement_1):S77–S93, 2020.

[LCLS10] Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, pages 661–670, 2010.

[LR85] Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.

[LWAL18] Haipeng Luo, Chen-Yu Wei, Alekh Agarwal, and John Langford. Efficient contextual bandits in non-stationary worlds. In *Conference On Learning Theory*, pages 1739–1776. PMLR, 2018.

[MNSDG09] M Teresa Melo, Stefan Nickel, and Francisco Saldanha-Da-Gama. Facility location and supply chain management–a review. *European journal of operational research*, 196(2):401–412, 2009.

[MZ21] A. Mills and S. Ziya. Testing with limited capacity and pooling. *MedRN: Respiratory Tract Infections (Topic)*, 2021.

[PRC+16] Vianney Perchet, Philippe Rigollet, Sylvain Chassang, Erik Snowberg, et al. Batched bandit problems. *Annals of Statistics*, 44(2):660–681, 2016.

[RDJ20] Kamalini Ramdas, Ara Darzi, and Sanjay Jain. 'test, re-test, re-test': using inaccurate tests to greatly increase the accuracy of covid-19 testing. *Nature medicine*, 26(6):810–811, 2020.

[Tho33] William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

[TM17] Ambuj Tewari and Susan A Murphy. From ads to interventions: Contextual bandits in mobile health. In *Mobile Health*, pages 495–517. Springer, 2017.

[VCL+20] Claire Vernade, Alexandra Carpentier, Tor Lattimore, Giovanni Zappella, Beyza Ermis, and Michael Brueckner. Linear bandits with stochastic delayed feedback. In *International Conference on Machine Learning*, pages 9712–9721. PMLR, 2020.

[Wor20] World Travel and Tourism Council. Recovery scenarios 2020 & economic impact from covid-19. `https://wttc.org/Research/Economic-Impact/Recovery-Scenarios`, November 2020.

Appendix A: Designing the Testing Supply Chain: Mathematical Optimization Formulation

In this section, we describe our stochastic, mixed-binary linear optimization formulation for designing the testing supply chain. Let $e \in \mathcal{E}$ index points of entry, and let $\bar{A}(e)$ denote the projected number of arrivals at e.

We first begin by modeling the expected number of infections caught at e under targeted testing when we allocate $b(e)$ tests. In an ideal world (without the need for exploration), Eva would use PLF data to predict the probability that a given passenger is COVID-19 positive before arrival. Eva would then order passengers in decreasing order of these probabilities and choose to test the top $b(e)$ passengers. The expected number of infections caught is then

$$\begin{aligned} \max_{\boldsymbol{z}} \quad & \sum_{i=1}^{\bar{A}(e)} \mathbb{P}(\text{Passenger } i \text{ is infected}) z_i \\ \text{s.t.} \quad & \sum_{i=1}^{\bar{A}(e)} z_i \le b(e), \\ & 0 \le z_i \le 1, \qquad i = 1, \dots, \bar{A}(e). \end{aligned}$$

By Lagrangian duality, this optimization is equivalent to

$$\min_{\theta \ge 0} \quad \theta b(e) + \sum_{i=1}^{\bar{A}(e)} \left[\mathbb{P}(\text{Passenger } i \text{ is infected}) - \theta \right]^+.$$

Letting R_e be a random variable denoting the probability that a *randomly* chosen passenger at e is positive, we can rewrite the above optimization as $\bar{A}(e) \cdot h\left(\frac{b(e)}{\bar{A}(e)}, e\right)$, where

$$h(p, e) \equiv \min_{\theta \ge 0} \quad \theta p + \mathbb{E}\left[(R_e - \theta)^+ \right]$$

This representation makes clear that $h(\cdot, e)$ is concave; for a fixed e it is the minimum of an (infinite) number of linear functions indexed by θ. Notice, p has the simple interpretation as the fraction of arrivals tested at port e and the functions $h(\cdot, e)$ can be pre-computed to arbitrary accuracy over the domain $p \in [0, 1]$.

Equipped with this function, we can state our formulation. Let $l \in \mathcal{L}$ index laboratories. For each $l \in \mathcal{L}$, let $c(l)$ denote lab l's daily processing capacity. Recall, not all labs may serve all points of entry due to geographical/logistical constraints. Let $\mathcal{C} \subseteq \mathcal{E} \times \mathcal{L}$ denote the set of compatible pairs of labs and points of entry. Moreover, let $\mathcal{L}_{wet}$ and $\mathcal{L}_{dry}$ denote the set of labs capable of processing wet and dry swabs respectively.

Our primary decision variables are $t_{wet}(e, l)$ and $t_{dry}(e, l)$ denoting, respectively, the number of wet (resp. dry) swabs processed by lab l from port of entry e. For convenience, we also introduce auxiliary variables $b(e)$ representing the total number of tests processed from port of entry e, the auxiliary (binary) variable $f(e, l)$ indicating whether lab l processes *any* tests from port of entry e, the auxiliary (binary) variables $g_{dry}(e)$, $g_{wet}(e)$ encoding if entry e uses dry and wet swabs respectively, and the binary variables $g(e)$ encoding if entry e uses both types of swabs.

We then solve

$$
\begin{aligned}
\max \quad & \sum_{e\in\mathcal{E}} \bar{A}(e)x(e) - \beta \sum_{(e,l)\in\mathcal{C}} f(e,l) - \gamma \sum_{e\in\mathcal{E},} g(e) \\
\text{s.t.} \quad & x(e) \;\le\; h\left(\frac{b(e)}{\bar{A}(e)}, e\right), && \forall e\in\mathcal{E}, && \text{(2a)}\\
& b(e) \;=\; \sum_{l\in\mathcal{L}} t_{wet}(e,l) + t_{dry}(e,l), && \forall e\in\mathcal{E} && \text{(2b)}\\
& b(e) \;\le\; \bar{A}(e), && \forall e\in\mathcal{E}, && \text{(2c)}\\
& \sum_{e\in\mathcal{E}} t_{dry}(e,l) + t_{wet}(e,l) \;\le\; 0.8c(l), && \forall l\in\mathcal{L}, && \text{(2d)}\\
& t_{wet}(e,l) \;\le\; f(e,l)\bar{A}(e), \;\; t_{dry}(e,l) \;\le\; f(e,l)\bar{A}(e) && \forall (e,l)\in\mathcal{C}, && \text{(2e)}\\
& t_{dry}(e,l) \;\le\; \bar{A}(e)g_{dry}(e), \;\; t_{wet}(e,l) \;\le\; \bar{A}(e)g_{wet}(e), && \forall (e,l)\in\mathcal{C}, && \text{(2f)}\\
& g_{dry}(e) + g_{wet}(e) \le 1 + g(e), && \forall e\in\mathcal{E}, && \text{(2g)}\\
& t_{dry}(e,l) \;=\; 0, && \forall l\notin\mathcal{L}_{dry}, && \text{(2h)}\\
& t_{wet}(e,l) \;=\; 0, && \forall l\notin\mathcal{L}_{wet}, && \text{(2i)}\\
& t_{dry}(e,l), t_{wet}(e,l), b(e) \ge 0, \;\; f(e,l), g(e)\in\{0,1\} && \forall (e,l)\in\mathcal{C}.
\end{aligned}
$$

At optimality, constraints (2a) will all be tight, and, hence, the three terms in the objective function correspond to i) the expected number of infections caught by targeted testing, ii) shipment costs, and iii) operational costs associated to storing both types of swabs. Here, $\beta \ge 0$ is a parameter (which we vary) which encodes the relative benefit of identifying an extra infection versus paying an additional fixed-cost for shipment. Similarly, $\gamma \ge 0$ is a parameter (which we vary) which encodes the relative benefit of identifying an extra infection versus the complexity of storing both types of swabs at a port. As mentioned in the main text, precisely estimating these constants is impossible, and our decision-support tool allowed policy-makers to inspect solutions for any choice of β and γ to choose a reasonable supply chain.

Constraint (2b) ensures $b(e)$ correctly represents the total tests available at entry e, while constraint (2c) ensures we do not allocate more tests than expected arrivals. Constraint (2d) ensures no more than 80% of the testing capacity at lab l is utilized. The parameter 80% was chosen exogenously based on intuition from queuing theory that some slack is necessary to ensure fast response times and robustness to disruptions.

Constraints (2e) force the binary variable $f(e,l)$ to be 1 if any tests at e are served by l. Similarly, constraints (2f) and (2g) ensure the binary variable $g(e)$ is 1 if both swab types are used.

Finally, since we only define the decision variables $t_{dry}(e,l)$ and $t_{wet}(e,l)$ for pairs $(e,l)\in\mathcal{C}$, the constraints (2h) and (2i) ensure that each entry routes tests to a compatible lab given their swab types.

Because of the concavity of $h(\cdot,e)$, constraint Eq. (2a) is non-linear, and one cannot pass the above formulation directly to a linear optimization solver. One remedy is to use the explicit representation of $h(\cdot,e)$ to replace constraint (2a) by the family of constraints,

$$x(e) \;\le\; \theta\frac{b(e)}{\bar{A}(e)} + \mathbb{E}\left[(R_e-\theta)^+\right] \qquad \forall\theta\ge 0,\ \forall e\in\mathcal{E}.$$

One can then pass the resulting formulation to a linear optimization solver and use constraint generation/lazy-cuts to sequentially add constraints corresponding to violated values of θ.

In practice, recognizing that $h(\cdot, e)$ is noisily estimated, we take a simpler approach and approximate $h(\cdot, e)$ by a piecewise-linear, concave function, i.e.,

$$h(p, e) \approx \min_{k=1,\dots K} h_1(e,k)p + h_0(e,k),$$

offline. We then replace constraint Eq. (2a) by the (finite) number of constraints

$$x(e) \;\le\; h_1(e,k)\frac{b(e)}{\bar{A}(e)} + h_0(e,k), \quad \forall k = 1,\dots,k,\ \forall e \in \mathcal{E}.$$

As shown in the main text, even for K as small as 3, the fit is exceedingly good.

In our implementation, we estimated the expectation defining $h(\cdot, e)$ by sample average approximation using data from a pilot roll-out of EVA preceding deployment.

Appendix B: Details of Empirical Bayes Procedure

In this section we provide details on our empirical Bayes estimation procedure. We also refer the interested reader to (BDG$^+$21) which contains some additional (minor) details on how this procedure was implemented with respect to countries with different risk tiers (also known as color designations).

We assume a set of $\mathcal{K}_t$ types available at time t. For each type k, we let $n(t,k)$ denote the number of tests performed on type k passengers in the past two weeks, and $X(t,k)$ the number positives identified. For simplicity, assume $n(t,k) > 1$ for all k. Let $R(t,k)$ denote the true, unknown prevalence of type k at time t.

We posit the following statistical model:

$$\begin{aligned} R(t,k) &\sim \text{Beta}(\alpha,\beta) \quad k \in \mathcal{K}_t, \\ X(t,k)|R(t,k), n(t,k) &\sim \text{Binom}(n(t,k), R(t,k)) \quad \forall k \in \mathcal{K}_t. \end{aligned} \tag{3}$$

In words, all prevalences are drawn from a common Beta distribution (the prior), and passengers test positively independently with probability equal to the prevalence of their type. The parameters (α, β) are unknown constants.

B.1. Variance of the MLE $\hat{r}^{MLE}(t,k)$

Equipped with the above notation, the maximum likelihood estimator is

$$\hat{r}^{MLE}(t,k) \equiv \frac{X(t,k)}{n(t,k)}.$$

Recall that the conditional standard deviation of this estimator is

$$\text{Std Dev}\left(\hat{r}^{MLE}(t,k) \,|\, R(t,k)\right) \;=\; \sqrt{\frac{R(t,k)(1-R(t,k))}{n(t,k)}}$$

For rarely occurring types, $n(t,k)$ may be less than 100, while typical prevalence is only 2/1000. In these settings, we might expect that

$$\frac{1-R(t,k)}{n(t,k)} \;\approx\; \frac{1}{n(t,k)} \;\ge\; R(t,k),$$

suggesting the above standard deviation is larger than the estimate itself. Hence, for such types, any signal is washed out completely by noise.

To provide some additional intuition, recall from the discussion preceding Fig. 7 that, using the two weeks of data preceding Sept. 1, 2020, the excess MSE of $\hat{r}^{MLE}(t,k)$ over a baseline estimator that always predicts 0 is approximately 0.000334. Hence, we have that

$$\begin{aligned}\sqrt{\frac{1}{|\mathcal{K}_t|}\sum_{k\in\mathcal{K}_t}\text{MSE}\left(\hat{r}^{MLE}(t,k)\right)} &\geq \sqrt{\frac{1}{|\mathcal{K}_t|}\sum_{k\in\mathcal{K}_t}\text{Excess MSE}\left(\hat{r}^{MLE}(t,k)\right)}\\ &= \sqrt{0.000334}\\ &\approx 0.0183,\end{aligned}$$

as claimed in the main body.

B.2. Moment Matching Equations for Empirical Bayes

Consider the (unconditional) first two moments of the $R(t,k)$. A straightforward computation shows that

$$M_1 \equiv \frac{1}{|\mathcal{K}_t|}\sum_{k\in K_t}\mathbb{E}\left[\frac{X(t,k)}{n(t,k)}\right] = \frac{1}{|\mathcal{K}_t|}\sum_{k\in K_t}\mathbb{E}\left[R(t,k)\right] = \frac{\alpha}{\alpha+\beta}. \tag{4}$$

$$M_2 \equiv \frac{1}{|\mathcal{K}_t|}\sum_{k\in K_t}\mathbb{E}\left[\frac{X(t,k)\big(X(t,k)-1\big)}{n(t,k)\big(n(t,k)-1\big)}\right] = \frac{1}{|\mathcal{K}_t|}\sum_{k\in K_t}\mathbb{E}\left[R(t,k)^2\right] = \frac{\alpha^2}{(\alpha+\beta)^2}+\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}. \tag{5}$$

In our procedure, we estimate these moments by the sample moments, yielding the following estimating equations:

$$\hat{M}_1 \equiv \frac{1}{|\mathcal{K}_t|}\sum_{k\in K_t}\frac{X(t,k)}{n(t,k)} = \frac{\hat{\alpha}}{\hat{\alpha}+\hat{\beta}} \tag{6}$$

$$\hat{M}_2 \equiv \frac{1}{|\mathcal{K}_t|}\sum_{k\in K_t}\frac{X(t,k)\big(X(t,k)-1\big)}{n(t,k)\big(n(t,k)-1\big)} = \frac{\hat{\alpha}^2}{(\hat{\alpha}+\hat{\beta})^2}+\frac{\hat{\alpha}\hat{\beta}}{(\hat{\alpha}+\hat{\beta})^2(\hat{\alpha}+\hat{\beta}+1)}. \tag{7}$$

We can solve these two equations for the two unknowns $(\hat{\alpha},\hat{\beta})$. The arithmetic is somewhat easier if we instead work in the parameterization $(\hat{r}^{prior},\hat{S}_0)$ where $\hat{r}^{prior}=\frac{\hat{\alpha}}{\hat{\alpha}+\hat{\beta}}$ and $\hat{S}_0=\hat{\alpha}+\hat{\beta}$. Then,

$$\hat{r}^{prior} = \frac{1}{|\mathcal{K}_t|}\sum_{k\in K_t}\frac{X(t,k)}{n(t,k)} \tag{8}$$

$$\hat{S}_0 = \frac{\hat{r}^{prior}-\hat{M}_2}{\hat{M}_2-(\hat{r}^{prior})^2}. \tag{9}$$

Plugging these estimates into Eq. (1) yields our prevalence estimates.

As a technical aside, these estimators are only well-defined when $\hat{M}_2-(\hat{r}^{prior})^2>0$. Inspection shows the left side of this inequality converges almost surely as $|\mathcal{K}_t|\to\infty$ to the variance of $R(t,k)$, and hence should be positive for $|\mathcal{K}_t|$ sufficiently large.

We next show that our estimates are Bayes-optimal and enjoy perturbation monotonicity when the number of types is large.

PROPOSITION 1 **(Optimality and Perturbation Monotonicity)**. *Suppose the data are drawn from the model Eq.* (3) *with* $\alpha,\beta>0$.

i) For each k, $\hat{r}^{EB}(t,k)=r^{Bayes}(t,k)+\mathcal{O}_p\left(\frac{1}{\sqrt{|\mathcal{K}_t|}}\right)$, *even if* $n(t,k)$ *is fixed as* $|\mathcal{K}_t|\to\infty$. *In other words, our estimates approach the Bayes optimal estimates as the number of types grows large, even if the amount of data per type is small.*

ii) When adding exactly 1 *positive case to the data, our EB estimates can decrease by at most* $\mathcal{O}_p\left(\frac{1}{\sqrt{|\mathcal{K}_t|}}\right)$. *In other words, our EB estimates satisfy perturbation monotonicity asymptotically as the the number of types grows large.*

Proof. First, observe by the central limit theorem that $\hat{M}_1 = M_1 + \mathcal{O}_p\left(\frac{1}{\sqrt{|\mathcal{K}_t|}}\right)$ and $\hat{M}_2 = \mathcal{O}_p\left(\frac{1}{\sqrt{|\mathcal{K}_t|}}\right)$ as $|\mathcal{K}_t| \to \infty$, even if all $n(t,k)$ are bounded by constant. Consequently, by the delta method and a Taylor Series expansion, we have that $\hat{r}^{prior} = r^{prior} + \mathcal{O}_p\left(\frac{1}{\sqrt{|\mathcal{K}_t|}}\right)$ and $\hat{S}_0 = S_0 + \mathcal{O}_p\left(\frac{1}{\sqrt{|\mathcal{K}_t|}}\right)$. Since $n(t,k) \geq 2$ for all k, our estimates $\hat{r}^{EB}(t,k)$ are differentiable functions of $(\hat{r}^{prior}, \hat{S}_0)$. Hence, by another application of the delta method, we have that $\hat{r}^{EB}(t,k) = r^{Bayes}(t,k) + \mathcal{O}_p\left(\frac{1}{\sqrt{|\mathcal{K}_t|}}\right)$, where $r^{Bayes}(t,k)$ is the optimal Bayes estimator (c.f. Eq. (1)) which depends on the (unknown) α, β. This proves the first claim.

To prove the second claim, it suffices to show that the Bayes estimator satisfies perturbation monotonicity. However, this follows immediately from Eq. (1) since the only data-dependent quantity is $\hat{r}^{MLE}(t,k)$ which remains unchanged when $k \neq \ell$ and only increases for $k = l$. □

Appendix C: Details of Test Allocation Algorithm

To keep the paper self-contained, this appendix briefly summarizes the key steps of our allocation algorithm from (BDG$^+$21). We refer the reader to that partner paper for additional details.

Recall that our estimation procedure yields a Beta posterior distribution. Let (α_k,β_k) denote the parameters of our posterior for type k. Let $F_{\alpha,\beta}$ be the CDF of a Beta distribution with parameters α and β. Then, from (GF16), 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,

$$\lambda_k = \frac{\alpha_k}{\alpha_k + \beta_k}\left(1 - \gamma F_{\alpha_k+1,\beta_k}(\lambda_k)\right) + \gamma\lambda_k\left(1 - F_{\alpha_k,\beta_k}(\lambda_k)\right), \tag{10}$$

which can be solved by bisection.

As discussed in Sec. 5.1, we perform certainty-equivalent updates to capture the likely reduction in the variance of our posterior distributions. Specifically, after allocating a single additional test to type k, we expect to observe a positive result with probability $\hat{r}^{EB}(t,k)$ and a negative result with probability 1-$\hat{r}^{EB}(t,k)$ in at most 48 hours. Consequently, after allocating a test, we update the parameters of our Beta distribution with this expected result, i.e., $\hat{\alpha}_k \leftarrow \hat{\alpha}_k + \hat{r}^{EB}(t,k)$, and $\hat{\beta}_k \leftarrow \hat{\beta}_k + 1 - \hat{r}^{EB}(t,k)$. Note that this CE update does not change our estimate of the mean prevalence, but it does reduce the variance of our posterior distributions.

Let $A(t,k,e)$ denote the number of scheduled type k arrivals at time t at port e. Further, let our testing allocations of type k passengers on day t at port of entry e be denoted by $n(t,k,e)$. These allocations must additionally satisfy constraints on testing budgets and arrivals. Specifically,

$$\sum_{k\in\mathcal{K}} n(t,k,e) \leq b_e \quad \forall e \in \mathcal{E}, \tag{11}$$

ensuring that the total number of allocated tests does not exceed the budget at each port of entry $e \in \mathcal{E}$, and

$$n(t,k,e) \leq A(t,k,e) \quad \forall e \in \mathcal{E}, \tag{12}$$

ensuring that the number of allocated tests for passengers of type k does not exceed the number of arriving type k passengers at every port of entry e.

Once our bandit algorithm identifies a passenger type to test, we must choose a specific port of entry at which to allocate the test; this decision affects where testing budget is consumed. We employ a simple greedy heuristic. Specifically, if type k currently has the highest optimistic gittins index, we choose a passenger of type k at the port of entry with the most remaining tests available. In other words, we preferentially allocate tests at less constrained ports, with the goal of potentially saving tests for rare types at constrained ports.

When a port of entry's budget is depleted, we remove that port of entry from consideration; similarly, if all arrivals of a certain type have been assigned to be tested, we remove that type from consideration.

The pseudocode for the resulting algorithm is given in Algorithm 1 below.

Algorithm 1 Test Allocation

Input: Posterior distribution estimates $\left\{\hat{\alpha}_k, \hat{\beta}_k\right\}_{k \in K_t}$ for each type k, arrivals $A(t,k,e)$ and budgets b_e for each type k and port of entry e, number of pipeline tests Q_k, discount factor γ
for $k \in \mathcal{K}_t$
    Compute λ_k from Eq. (10) using $\hat{\alpha}_k,\ \hat{\beta}_k$
    $\lambda_k \leftarrow$ CE update index of type k (repeat Q_k times) # *account for pipeline tests*
    $Y_k \leftarrow \sum_e A(t,k,e)$ # *type k passengers not yet allocated a test*
    for $e \in \mathcal{E}$
        $Y_{ke} \leftarrow A(t,k,e)$ # *type k passengers at port of entry e not yet allocated a test*
        $C_e \leftarrow b_e$ # *remaining (un-allocated) tests at port of entry e*
        $n(t,k,e) \leftarrow 0$ # *initialize allocations*
    end
end
while $\max_e C_e > 0$ **and** $\max_{k,e' \in \{e|C_e>0\}} Y_{ke'} > 0$
    $k^* = \arg\max \{\lambda_k : Y_k > 0\}$
    $e^* = \arg\max \{C_e : Y_{k^*e} > 0\}$
    $n(t,k^*,e^*) \leftarrow n(t,k^*,e^*) + 1$ # *allocate a test to a type k passenger at port of entry e*
    $C_{e^*} \leftarrow C_{e^*} - 1, Y_{k^*e^*} \leftarrow Y_{k^*e^*} - 1, Y_{k^*} \leftarrow Y_{k^*} - 1$
    $\lambda_{k^*} \leftarrow$ CE update index of type k^*
end
return $n(t,k,e)_{k \in K_t, e \in \{1,\ldots,\mathcal{E}\}}$ # *number of tests allocated to each type at each port of entry*