Reinforcement Learning for Public Health: Targeted COVID-19 Screening

Hamsa Bastani[1][0000−0002−8793−4732], Kimon Drakopoulos[2][0000−0001−8288−5874], and Vishal Gupta[2][0000−0003−4371−9114]

[1] University of Pennsylvania, Philadelphia PA 19104, USA
hamsab@wharton.upenn.edu
[2] University of Southern California, Los Angeles CA 90089, USA
{drakopou,guptavis}@usc.edu

Abstract. Reinforcement learning is a promising solution to sequential decision-making problems, but its use has largely been limited to simulation environments and e-commerce. This chapter describes a large-scale deployment of reinforcement learning in Greece during the summer of 2020 to adaptively allocate scarce testing resources to incoming passengers amidst the evolving COVID-19 pandemic. Our system, nicknamed Eva, used limited demographic information and recent testing results to guide testing in order to maximize the number of asymptomatic but infected travelers identified over the course of the tourist season. Results from the field evaluation show a marked improvement over other "open-loop" testing strategies and highlight some of the challenges of deploying reinforcement learning in real-world, high-stakes settings.

Keywords: contextual bandits · empirical Bayes · disease screening.

This chapter summarizes our deployment of a reinforcement learning algorithm for targeted COVID-19 border testing in Greece; further background and technical details can be found in our journal articles [1] and [2].[3]

1 Introduction

In the beginning of 2020, governments around the world scrambled to mount an emergency response to the COVID-19 pandemic. At the time, vaccines for SARS-CoV-2 were unavailable and effective medical treatments for COVID-19 were still in development. Testing, contact tracing and social distancing were the only weapons against the pandemic. Hence, this global response focused on a largely uncoordinated flurry of lockdowns, enhanced social distancing protocols, and travel restrictions.

[3] Anonymized and deidentified data can be found in https://github.com/kimondr/EVA_Public_Data; code for our counterfactual analysis can be found in https://github.com/vgupta1/EvaTargetedCovid19Testing.

However, as became quickly apparent, these restrictive measures could not be sustained indefinitely as they effectively halted normal economic activity. For European countries with significant tourist economies, like Greece, travel restrictions were especially challenging, with projected losses of $1 trillion USD and 19 million jobs in the tourist industry alone [3].

Consequently, as the first tourist season of the pandemic approached (summer of 2020), many countries sought to judiciously relax travel restrictions. Striking the right balance was crucial. Too loose travel protocols risked admitting asymptomatic or presymptomatic infected travelers who might then spread SARS-CoV-2 among the local population with devastating consequences. Too strict protocols would inhibit the desired economic activity.

An ideal solution might have been to actively screen *every* passenger arriving at the border for SARS-CoV-2. Positive cases could then be quarantined, preventing them from spreading SARS-CoV-2 to the local population on their visit. However, in early 2020, testing supply chains across the globe were strained. Indeed, Polymerase Chain Reaction (PCR) testing requires specialized machinery, and the backlog on these machines was 3-6 months, precluding rapid capacity expansion. Although other forms of antigen testing were more readily available, their accuracy, particularly among asymptomatic/presymptomatic passengers, was unclear. In summary, we simply could not test *everyone.* Accurate testing was a scarce resource that needed to be managed carefully.

Against this backdrop, we partnered with Greece to design and deploy Eva, a reinforcement learning system that combined passenger demographic information with recent testing results from similar passengers to allocate tests at the Greek border. The key idea was to preferentially test travelers who, based on recent data, were *more likely* to test positive. The key challenge was that COVID-19 prevalence was constantly evolving as the pandemic ebbed and flowed. Risk across various subpopulations might change substantively over time, with previously "safe" populations becoming high risk due to a litany of factors such as eased social distancing or individual behavioral modifications. Hence, any good testing strategy should be adaptive, reacting to the changing nature of the pandemic.

The reader might recognize this description as the classic "exploration – exploitation" tradeoff in sequential learning with bandit feedback [5] – we need to "exploit" by allocating tests primarily to high-risk individuals, but we must also "explore" by testing broadly to identify any previously "safe" populations that are becoming risky. At the heart of Eva is a contextual bandit algorithm that navigates this tradeoff, but, as we discuss below, our real-world setting introduces a number of unique challenges including highly-imbalanced data, time-varying dynamics, supply-chain constraints, and batched feedback.

1.1 Contrast to Conventional Travel Protocols

To the best of our knowledge, Eva was the first attempt at a purely data-driven protocol for screening and travel restrictions. To appreciate the value of reinforcement learning in this setting, it may help to contrast Eva to the travel protocols used by other countries in the European Union at the time.

Most national-level travel protocols proceeded by designating sets of countries of origin as white-, grey-, red-, or black-listed. Roughly speaking, travelers from white-listed countries would be allowed to enter freely; travelers from grey-listed countries would need to provide proof of a recent negative PCR test to enter; travelers from red-listed countries would need to quarantine for a significant period upon arrival; travelers from black-listed countries would be forbidden from entering for any reasons that were deemed non-essential. Importantly, different nations would differ in their color designations of origin countries, in part because these designations were largely informed by human judgement. More specifically, one would ideally use data on COVID-19 prevalence within the asymptomatic traveler population from an origin country to determine its color designation. However, at the time, such data was non-existent because testing was scarce. Consequently, decision-makers instead used proxy data in the form of publicly reported population-level epidemiological metrics (new cases per capita, new deaths per capita, and positivity rates in testing). The gap between proxy data and the ideal COVID-19 prevalence rates in traveler populations left some room for interpretation.

By contrast, Eva's reinforcement learning strategy sought to directly collect data on prevalence in the populations of interest as it was screening. Although this is clearly more costly (both financially and operationally) than using proxy data, as we will discuss later on, the higher-quality data directly translates to superior performance. In essence, this is the key value proposition in favor of reinforcement learning for this setting.

2 Problem Statement

As discussed below, at least 24 hours prior to travel, each passenger submits a Passenger Locator Form which contains some basic demographic information (including their age, gender, and origin country) and projected itinerary in Greece. (See Sections 3 and 4 below for details on how this form was designed and what it contains.) Importantly, since all collected information is categorical, there are a finite number of possible traveler features $\mathcal{X}$.

Furthermore, there are approximately 40 entry points to Greece, including land borders, seaports and airports. Denote the set of entry points by $\mathcal{E}$. Let $A_{x,e}(t)$ denote the number of arrivals of travelers with feature x at entry e on day t. Finally, let $B_e(t)$ denote an entry-specific testing budget determined exogenously by the Secretary General of Public Health. (See also Section 3 below for details on how these budgets were determined.)

In addition to these (known) operational quantities, let $R_x(t)$ denote the *unknown* underlying risk for a passenger with feature x on day t. In particular, a random passenger with features x tested on day t will test positive with probability $R_x(t)$.

Our goal is to determine the number of tests $T_{x,e}(t)$ to allocate to passengers with feature vector $x \in \mathcal{X}$, arriving at entry e on day t in order to maximize the expected number of infections caught at the border over the course of the entire time horizon $\mathcal{T}$ (summer of 2020) subject to various testing constraints. Mathematically, we seek to maximize

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

subject to entry-specific budget constraints on testing,

$$\sum_{x\in\mathcal{X}} T_{x,e}(t) \le B_e(t) \text{ for every entry } e \in \mathcal{E} \, ,$$

and the constraint that we cannot test more travelers than actually arrive at a particular entry on a particular day,

$$T_{x,e}(t) \le A_{x,e}(t) \text{ for every feature } x \in \mathcal{X} \text{ and entry } e \in \mathcal{E} \, .$$

Importantly, our decisions are $T_{x,e}(t)$ are adapted, i.e., they can depend on any data available before time t. Finally, we stress that, we do not know the true traveler risk $R_x(t)$. Instead, we must estimate $\hat{R}_x(t)$ using testing results from previous time steps, resulting in an exploration-exploitation tradeoff in the choice of $T_{x,e}(t)$.

3 Practical Considerations and Modeling

The previous description of the testing problem is stylized and abstracts away many of the practical constraints and real-world issues of deployment. We next provide a few indicative examples of such issues.

Determining the Feature Set $\mathcal{X}$. First, the above description of the problem assumes the set of available features is fixed and exogenously given. In reality, before the deployment of Eva, we worked with the Secretary General of Public Health, and Data Protection Officers to design the PLF which solicited these features. In other words, we could partially control the set $\mathcal{X}$.

Designing the PLF again entails a tradeoff: On the one hand, we would like as much epidemiologically relevant data as possible about each traveler (past medical records, their individual risky behaviors, their occupation, etc.). On the other hand, such detailed data is particularly invasive and raises ethical concerns. Striking a balance involved collaborating with a broad variety of experts including privacy scholars, lawyers, and epidemiologists, to i) identify a minimal set of informative data to solicit *and* ii) implement a transparent privacy policy about how that data would be used, how it would be protected, and how it could (upon passenger request) be deleted.

Designing the Testing Supply-Chain. Our previous discussion simply asserts the testing budgets $B_e(t)$ are specified exogenously and fixed, but where do these budgets come from? Some thought suggests that every test performed at a point of entry must be sent to a nearby laboratory to be processed. The budget $B_e(t)$ is thus affected by a host of supply-chain design choices including:

. The number of labs with which we choose to contract,
. To which labs we send samples from entry e,
. Individual Lab processing capacity as determined by available machines for processing PCR tests and staffing.

Hence, we worked closely with the Greek government to optimize this testing supply-chain and maximize our budget. As an example, we argued strongly for Dorfman group testing (group size 5) to increase the effective testing capacity without compromising accuracy. As a second example, we formulated a stochastic, mixed-binary linear optimization problem to assess different supply-chain configurations (i.e. determining which laboratories would serve which points of entry). Policymakers utilized outputs from this optimization formulation as a decision-support tool and find a suitable design that mitigated transportation and logistics costs while providing the adequate budget to all points of entries.

The Flow of Information. Finally, our stylized formulation in Section 2 obscures the precise evolution of information in the system. Specifically, because of concerns about internet connectivity at particularly remote points of entry, all testing decisions for the day needed to be made at the beginning of the day. This phenomenon is frequently termed “batching" in the bandit literature. Moreover, even after a passenger was tested, results would only be available $24 - 48$ hours later due to (random) lab processing times. This phenomenon is frequently termed “delayed-feedback" in the bandit literature. Consequently, when assigning a test to a passenger, we do not know the testing results of passengers previously tested on the same day, know only some of the results from passengers tested between yesterday and the day before, and know all results for passengers tested more than 48 hours previous.

This flow of information imposes a number of additional constraints on our system. For example, since decisions are batched, computed centrally and then transmitted to points of entry at beginning of the day, our algorithm must be exceedingly computationally efficient. Indeed, we needed to reliably compute all testing decisions each day in less than 5 minutes (see also Section 5.) Moreover, it is not enough to determine $T_{xe}(t)$, the number of x-passengers to test at entry e on day t, one must actually identify $T_{xe}(t)$ specific passengers from the list of arriving passengers to test, and identify them as they disembark. Finally, more generally, our reinforcement learning algorithm should handle the above information delays, as well as delays caused by any unexpected disruptions at individual laboratories. In particular, we would not want all tests for x-passengers performed at the same entry point or processed at the same lab, since an unexpected disruption might leave us with no information about x-passengers at all.

The above list of real-world considerations is far from exhaustive. (We refer the interested reader to [2] for more details.) In mentioning them, we only wish to stress that in any large-scale, real-world deployment of a reinforcement learning tool, the actual reinforcement learning algorithm is only one piece of a larger puzzle. There are often a host of surrounding operational and logistical challenges that need to be solved. Indeed, reinforcement learning only works well if there is a reliable, fast feedback loop of information, and solving the operational problems to facilitate that loop should be a first order concern.

4 Method

Figure 1 illustrates the operational flow of Eva.

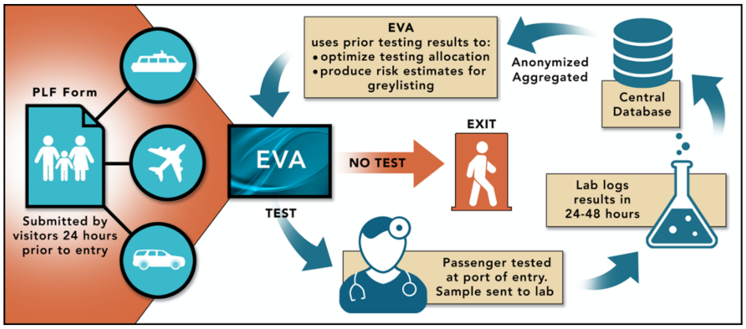

PLF Form
Submitted by visitors 24 hours prior to entry
EVA
EVA
uses prior testing results to:
• optimize testing allocation
• produce risk estimates for greylisting
Anonymized
Aggregated
Central Database
EXIT
NO TEST
TEST
Lab logs results in 24-48 hours
Passenger tested at port of entry. Sample sent to lab

Fig. 1: **Project Eva.** A reinforcement learning system for COVID-19 screening that we deployed during the summer of 2020 (Aug 6 to Nov 1) across all 40 points of entry to Greece, including airports, sea ports and land crossings. Figure reproduced from [1].

Information Collection: First, all travelers complete a PLF (one per household) at least 24 hours prior to arrival in Greece. This form serves to collect important information on features (origin region and country, age group, gender), operations (point and date of entry) and potential contact tracing (contact information, destination within Greece).

Risk Estimation: Next, we use recent testing results to form risk estimates of arriving travelers based on their feature information. Estimation is especially challenging since (i) our outcomes are highly imbalanced, since only 1-2 out of every 1000 travelers tested positive over the course of the summer, (ii) our training data is limited to a 14-day rolling window of the most recent testing results to ensure that our estimates are current, (iii) our primary predictive feature (origin location) is very high-dimensional since it is a categorical variable with a very large number of levels (e.g., we observed travelers from over 17,000 regions of origin).

We address these challenges in two steps. First, we use LASSO logistic regression [9] to reduce the dimensionality of the problem, grouping travelers into a small number of discrete types. Second, we use empirical Bayes [10] to learn a shared prior across different passenger types; this step is critical to reduce the variance of traveler types with very few arrivals. Figure 2 shows the resulting (anonymized) actual risk estimates on a specific day; this plot was featured on dashboards for Greek policymakers to inform domestic downstream decisions. [1] describes the technical specifications and associated details.

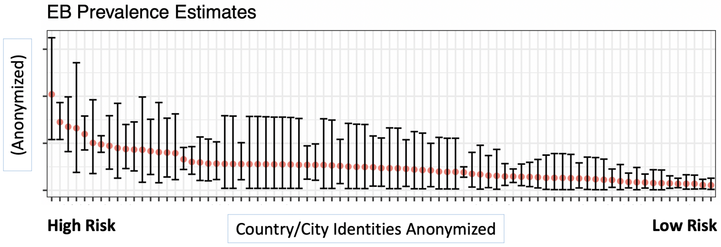


Fig. 2: **Risk Estimates.** Shown for each (low-dimensional) traveler type on a given day, ordered from high to low risk. We depict the mean and [5%, 95%] confidence intervals of the estimated posterior distribution. Figure reproduced from [1].

Allocating Tests: Equipped with high-quality risk estimates, we build on the contextual bandit literature [5] to target a subset of travelers for PCR testing based on their feature information. Unlike the classical contextual bandit literature, our problem *simultaneously* suffers nonstationary rewards (due to the continuously evolving risk estimates), batched decision-making (all testing decisions must be made at the start of the day), delayed feedback (results of allocated tests are returned only after 48 hours), and constraints (entry-specific budgets and arrivals). We propose a modified contextual bandit algorithm with *certainty-equivalent updates* to seamlessly handles these characteristics. Figure 3 shows the resulting (anonymized) actual test allocations (relative to arrivals) on a specific day; this plot was featured on dashboards for Greek policymakers to inform domestic downstream decisions. [1] describes the technical specifications and associated details.

Closing the Loop: All allocated tests were processed in a laboratory and sent to Eva's database after a 48-hour delay. These results informed risk estimates and test allocations in the next time periods. This dynamic feedback loop was critical to ensuring Eva's *agility* — closely following the dynamic ebb and flow of traveler risk across the globe — making targeted testing decisions that effectively safeguarded public health.

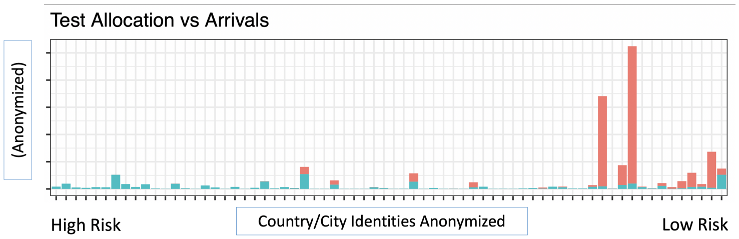


Fig. 3: **Test Allocations by Arrivals.** Shown for each (low-dimensional) traveler type on a given day, ordered from high to low risk. We depict the number of test allocations (teal) and untested scheduled arrivals (pink) for each type. Our algorithm allocates tests to essentially all high-risk arrivals (exploitation) but additionally assigns a small number of tests to all types (exploration); distortions are due to port-specific budget constraints. Figure reproduced from [1].

5 Resource Requirements

Eva was part of a larger system (travel.gov.gr) hosted by the Greek Ministry of Digital Governance that was built in three tiers: user interface, back-end and databases. The databases were hosted on Aurora PostgreSQL clusters and the back-end was designed in a Java-Quarkus framework. As with any national-scale ecosystem, there was significant heterogeneity in the access rights for different entities. Thus, each component of the system is its own Virtual Private Cloud (VPC), and the system only allows internal communication between components through peering, excluding communication with the external internet. For example, Eva was hosted on a separate VPC, and the resulting testing recommendations were sent to the internal email client of a different VPC that sent travelers their QR codes. Furthermore, Security Groups were defined to dictate access to information by different users. For example, the Eva VPC could only access aggregated and pseudonymized subsets of granular traveler data.

The system received and processed data from 30,000-100,000 households every day. Due to various operational constraints (discussed in Section 7), the allocation module of our system initiated at midnight, and all QR codes had to be sent to travelers within 5 minutes. As a result, computational demands were highly bursty. In order to ensure robustness while minimizing computational burden, we used load balancing (splitting job requests across many identical servers) and horizontal scalability (dynamically adjusting the number of servers used based on the demand).

6 Field Evaluation

Eva was deployed across all ports of entry to Greece from Aug. 6th to Nov. 1st 2020. During the peak season (Aug. 6th to Oct. 1st) when most tourist arrivals occurred, Eva processed approximately 41,830 (±12,784) PLFs each day and tested approximately 16.7% (±4.8%) of arriving households. At the conclusion of the summer, by simple counting, we knew exactly many positive cases were identified among those tested.

However, to assess efficacy, one must answer a counterfactual question: how many positive cases might we have identified with an alternate, simpler testing strategy? Indeed, if we could have identified a comparable number of cases, we might have attained similar public health benefits at a fraction of the cost.

In general, answering counterfactual questions is notoriously difficult. Since our alternate policy would have tested different passengers, how can we assess if those passengers would be positive? Fortunately, in our setting, we can leverage established techniques from off-policy evaluation, namely inverse propensity weight scoring, to estimate the number of positive cases the alternate testing strategy would have found. We can then compare this estimated number of positive cases identified under the alternate strategy to the actual number of positive cases found to assess effectiveness.

Using this strategy, we first compare Eva's performance to randomized, surveillance testing (see left panel of Figure 4). Recall, randomized surveillance testing was the first proposed approach by Greece. We calibrate the alternate strategy to use the same number of tests per port of entry as Eva, ensuring an "apples-to-apples" comparison. We find that during the peak tourist season, random surveillance testing would have identified 54.1% (±8.7%) of the number of positive cases as Eva.[4] We refer the interested reader to [1] for details on the calculation. Said differently, randomized surveillance testing would require $1/.547 \approx 1.85$ *times* as many tests as Eva required to achieve the same public health benefit, a substantive and expensive capital investment.

In the off-peak season (Oct.1 - Nov.6), randomized surveillance performs somewhat better (equivalently, Eva exhibits less of a benefit). We estimate that randomized surveillance would have identified 73.5% (±11.0%) of the cases that Eva did. We attribute this difference to increased availability of testing. Specifically, in the off-peak season, fewer tourists arrived, so it was possible to test a larger fraction of arrivals with the same testing budget. As the fraction of arrivals tested increased, the value of smart, targeted testing decreases. See right panel of Figure 4.

Of course, randomized surveillance testing is a weak benchmark. Throughout the summer, many countries published epidemiological data including the number of new cases, number of deaths, and positivity rate in testing. These data were far from perfect; many countries reported these figures sporadically, or would substantially revise published figures after a few days or weeks. Some countries reported nothing at all. Nontheless, these noisy data represented *some* signal that might be exploited, and a stronger benchmark might be based on this signal. Indeed, initial proposals in the European Union suggested basing travel protocols upon such published metrics [6, 7].

[4] For anonymity, we present statistics relative to the actual number of cases caught by Eva.

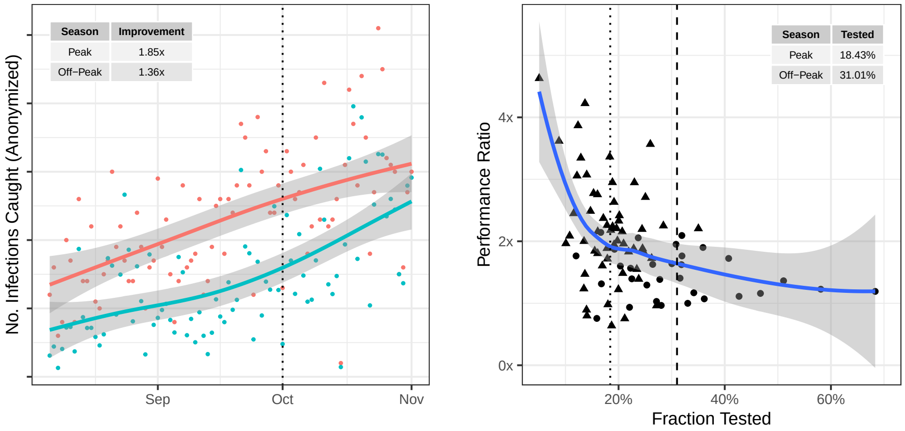
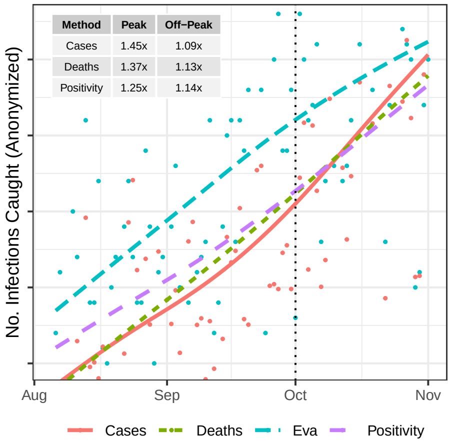


Fig. 4: **Comparing Performance of Eva vs. Randomized Testing.** Left Panel: We see the actual number of infections identified by Eva and the estimated number of infections identified by random, surveillance testing. Inset table gives the relative performance of Eva over the peak and off-peak tourist seasons. Right Panel: The daily performance ratio of Eva relative to random, surveillance testing as a function of the fraction of arrivals tested each day. Inset shows the average fraction of tested daily over peak and off-peak tourist season. Both figures reproduced from [1].

Consequently, we also compared the performance of Eva to 3 other benchmark policies corresponding to the 3 most popular epidemiological metrics: new cases per capita, new deaths per capita, and daily positivity rate. For each policy, the probability a passenger was tested was proportional to the current epidemiological metric of their country of origin. Hence, passengers from countries deemed more “risky” were tested with higher probability. Again, for an apples-to-apples comparison, we ensure that the number of tests performed at each point of entry matches those performed by Eva.

Figure 5 shows the comparison. Again, we see that Eva has a distinct edge over the benchmark policies. Specifically, policies based upon cases per capita, deaths per capita and positivity rates identified 69.0%(±9.4%), 72.7%(±10.6%), and 79.7%(±9.3%), respectively, of the infected travelers identified by Eva per test in the peak tourist season. Said differently, Eva identified 1.25x – 1.45x more infections with the same testing budget. In the off-peak season, this benefit drops to 1.09x - 1.15x. Again, we attribute the drop to increased availability of testing in the off-peak season.

Why do we see Eva consistently outperforming these benchmarks? While it is difficult to offer a definitive explanation, analysis in [1] strongly suggests that the issue is *not* that our benchmark policies were insufficiently exploiting the signal in the epidemiological data. Indeed, it seems *any* policy based on these data would underperform relative to Eva. Instead, we conjecture that the fundamental issue is that the measured epidemiological metrics provide signal



Method	Peak	Off−Peak
Cases	1.45x	1.09x
Deaths	1.37x	1.13x
Positivity	1.25x	1.14x

Fig. 5: **Comparing Performance of Eva vs. Benchmarks based on Epidemiological Metrics.** We compare the actual number of infections identified by Eva to the estimated number of infections identified by benchmark policies based on publicly reported cases per capita, deaths per capita, and positivity rate per country. Eva retains an advantage over such policies, with a larger benefit in the peak tourist season. Figure reproduced from [1].

about SARS-CoV-2 prevalence in the *general*, tested population of a country, while we are only interested in prevalence among the subset of the population that would travel to Greece in summer of 2020. Depending on the idiosyncratic biases in a specific origin country's testing protocol, the subset of would-be travelers is likely systematically younger, wealthier, and more risk-taking than the general, tested population. This discrepancy causes the signal in the reported epidemiological metrics to be only weakly informative. Indeed, this observation is perhaps the strongest argument for a reinforcement learning approach in this application – since it allows one to collect high-quality data directly on the population of interest.

7 Lessons learned

We next discuss some lessons learned from the design, development and deployment of Eva.

Privacy preservation is not just about hiding names. It is important to differentiate between deidentification and anonymization. Neither masking sensitive identifying information by assigning unique ids nor fully removing data fields with unique identifying information (de-identification) is equivalent, legally or ethically, to fully protecting against individual reidentification (anonymization). Indeed, the latter task is substantially harder.

Consider the following simple example: Suppose that we only retained information about a traveler's age group (0-9 years, 10-19 years, 20-29 years, etc.), gender, region of origin, date of arrival, and testing status. On the surface, these data are de-identified – they do not contain any traveler's name or unique ID. One might believe this de-identification is sufficient for anonymization.

However, suppose there was only one female traveler in her 30s visiting Greece on August 1st, 2020 originating from the Gozo region of Malta and that this traveler tested positive. If a malicious party had access to the relevant airline manifest, they would be able to infer that this specific passengers name and that they tested positive — despite the fact that her name was hidden in the testing records — since her other attributes are sufficiently informative. Thus, with a passenger manifest, she is re-identifiable.

While there are many guidelines for anonymizing data such as ensuring subgroups are sufficiently coarse with at least n members, these guidelines can be hard to ensure in reinforcement learning contexts. For example, in our setting, we cannot control how many passengers of each type $x \in \mathcal{X}$ arrive, a priori, and, hence, cannot ensure subgroups are sufficiently coarse. Moreover, insofar as algorithms for re-identification by pairing with an auxiliary dataset improve every day, ensuring anonymization is an increasingly challenging task.

These observations underly the importance of *data minimization* (that is, requesting the minimum required information for a task) as instructed by the General Data Protection Regulation (GDPR). When designing real-world AI tools, all stakeholders should seek to obtain and store a minimal subset of features (least invasive) with sufficient predictive signal. Of course, data minimization often entail sacrificing performance, but managing the trade-off between privacy and effectiveness needs to be an explicit discussion at the inception of a project.

Modular design for continuous adaptation. Large-scale systems must be flexible in incorporating changes in the environment/policy or stakeholder's requirements. Pandemic mitigation is an extreme example of the need for flexibility, since travel protocols, testing technologies, availability of testing and availability of vaccines were constantly changing. In response to this requirement, the modular design of Eva, disassociating type extraction, estimation and test allocation allowed for seamlessly incorporating new types (e.g. vaccinated or not) and different testing delays (e.g. rapid vs. PCR); see the end of [1] for a discussion.

Simplicity and transparency for trust. Public-facing projects that are designed by interdisciplinary teams to support decision-making must provide transparent reasoning. In this vein, the modeling and design choices underlying Eva were made not only for efficiency but also for conceptual simplicity; this allowed all parties to evaluate and trust the system. For example, our combination of LASSO and empirical Bayes transformed the high-dimensional traveler features into a small, discrete set of traveler types with interpretable confidence intervals that behaved intuitively. While not the most sophisticated statistical machinery, this approach was readily understandable by medical collaborators and the epidemiologists on the team.

Beyond fostering trust among policy-makers, outputs from an AI project are often used to qualitatively support downstream policy-making. For example, our estimates were used across ministries of the Greek Government to position mobile testing units, guide smart contact-tracing, and adjust social distancing measures. Given the variety of potentially non-technical policy-makers leveraging outputs of the system to inform their decision-making, clarity and transparency of outputs and reasoning are critical.

The algorithm and the technology is not enough. As mentioned earlier in Section 3, in large-scale systems like Eva, the algorithm and technology components are only a piece of the larger puzzle. One must consider practical constraints of the larger ecosystem in the design phase of the algorithm. Often, a simpler algorithm that dovetails well with practical consideration will fare far better than a state-of-the-art algorithm that is too inflexible for real-world idiosyncracies.

8 Conclusion

Artificial Intelligence is still in its early stages of adoption, and every day researchers find new ways to leverage these tools to address real-world challenges. As we as a community continue to deploy AI solutions, we should strive for indisciplinary teams with expertise *beyond* algorithmic design. “Build a big tent" – one large enough for humanists, computer scientists, operations researchers, and policy-makers – and then leverage their expertise. The real-world problems we face are multifaceted and complex. AI alone won’t solve them, but, partnered with other approaches, it can be a crucial tool.

Acknowledgements

The authors are very grateful to Jon Vlachogiannis, Christos Hadjichristodoulou, Gkikas Magiorkinis, Dimitrios Paraskevis, Pagona Lagiou, and Sotirios Tsiodras for their indispensable role in Eva’s design and deployment. The authors further 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 also contributed to Eva’s deployment.

References

1. Bastani, H., Drakopoulos, K., Gupta, V., Vlachogiannis, J., Hadjicristodoulou, C., Lagiou, P., Magiorkinis, G., Paraskevis, D., Tsiodras, S.: Efficient and Targeted COVID-19 Border Testing via Reinforcement Learning. Nature **599**(7883), 108–113 (2021)

2. Bastani, H., Drakopoulos, K., Gupta, V., Vlachogiannis, J., Hadjicristodoulou, C., Lagiou, P., Magiorkinis, G., Paraskevis, D., Tsiodras, S.: Interpretable Operations Research for High-Stakes Decisions: Designing the Greek COVID-19 Testing System. INFORMS Journal on Applied Analytics (Forthcoming)
3. World Travel and Tourism Council, `https://wttc.org/Research/Economic-Impact/Recovery-Scenarios`. Last accessed Nov 2020
4. Mullainathan, S., Obermeyer, Z.: Diagnosing physician error: A machine learning approach to low-value health care. The Quarterly Journal of Economics **137**(2), 679–727 (2022)
5. Thompson, W.: On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. Biometrika **25**(3-4), 285–294 (1933)
6. Council recommendation on the temporary restriction on non-essential travel into the EU and the possible lifting of such restriction. `https://www.consilium.europa.eu/media/47592/st_9208_2020_init_en.pdf`. Last accessed: 30 June 2020
7. 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, 12 October 2020
8. Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions: Tourism and transport in 2020 and beyond, European Commission, 13 May 2020
9. Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological) **58**(1), 267–288 (1996)
10. Devine, O., Louis, T., Halloran, E. Empirical Bayes methods for stabilizing incidence rates before mapping. Epidemiology 622–630 (1994)