Data-driven Adaptive Testing Resource Allocation Strategies for Real-time Monitoring of Infectious Diseases

Xin Zan

Department of Industrial and Systems Engineering, University of Florida, Gainesville, FL 32611

Infectious diseases have continued to be a major global public health threat and effective methods are in critical need to quickly detect disease outbreaks. However, in practice, limited testing availability that leads to insufficient data poses challenges to effective analysis and real-time monitoring of infectious diseases, especially at the early stage of a novel infectious disease such as coronavirus disease 2019 (COVID-19). To tackle this challenge, this work proposes adaptive allocation strategies to intelligently allocate limited testing resources among communities, by integrating nonstationary Multi-Armed Bandit (MAB) techniques on top of a physics-informed model which accounts for transmission dynamics and health disparity embedded in infectious diseases. A comprehensive simulation study based on the COVID-19 pandemic in North Central Florida are conducted to evaluate the proposed methodology, showing an overall robust and satisfactory performance.

Keywords: data-driven; infectious diseases; MAB; real-time monitoring; physics-informed; resource allocation.

1 Introduction

Over the past several decades, infectious diseases have continued to be a major global public health threat, contributing significantly to the escalating costs of health care, and remaining one of the leading causes of death worldwide. For example, COVID-19, only the most impactful pandemic we have experienced, has a profound impact on global health and economic due to its rapid spread and evolution across the globe. To relieve the ongoing global burden of infectious diseases, a surge of interest in data science methods has been witnessed for early detection of outbreaks and support of timely implementation of public health interventions to contain the fast infectious spread through communities at the early stage [1]. For real-time surveillance of infectious diseases, mass testing is key to tracking and understanding the infection progression using diagnostic tests, which allows health care practitioners to appropriately and timely identify and isolate positive cases to contain the further spread of infectious diseases. Amid the fast worldwide spread of these infectious diseases, however, limited testing availability that leads to insufficient testing data poses significant challenges across the globe in effective analysis and monitoring of a novel infectious disease such as COVID-19, especially at the

early stage. There are two major practical restrictions of testing availability: (i) the limited collection of testing samples due to the demand of intensive labor; (ii) the severely restricted viral testing capability due to the scarcity of reactants, laboratory testing infrastructure, and trained technicians. Subject to limited testing availability, although pooled testing strategies have emerged as a useful means to increase testing capacity by testing a pooled sample with one diagnostic test, the resulted test reporting delays and lowered testing sensitivity interfere with the effective real-time monitoring [2].

This work focuses on effectively allocating limited individual diagnostic tests for mass testing, the capacity of which yet falls far short of the need for containing infectious diseases during a surge. The resulted inadequate reliable testing data greatly limits the effective analysis of the infectious disease progression and appropriate health risk assessment in different communities, thus impeding the proper and prompt identification of disease outbreaks which will lead to delayed and erratic containment policies during the surveillance. To tackle these challenges, the primary contribution of this work is to maximally leverage all available physical information to develop data-driven adaptive testing resource allocation strategies which strategically and dynamically determine where and how many tests should be conducted at each time to collect high-quality testing data for quick detection of disease outbreaks in the context of limited testing availability. It should be noted that the proposed method can be flexibly applied to general pandemics due to its data-driven and physics-informed framework based on domain knowledge of general communicable infectious diseases.

2 Methodology development

In this section, we will present the proposed method subject to limited testing availability in detail. To highlight our main idea, several fundamental assumptions are made: (i) the diagnostic tests are allocated based on the geographical unit of census block group (BG). Denote the set of all BGs by G, the population for BG i by B_i, and the whole population by $\mathcal{B} = \sum_{i \in G} B_i$; (ii) the sampling within each BG is random; (iii) the total number of limited diagnostic tests is fixed over time, i.e., $Q = \sum_{i \in G} q_i(t), \forall t$. At each time t, according to the allocation strategy $\boldsymbol{q}(t)$, the collected mass testing results are the number of positive cases $n_i(t)$ for each BG. With the use of the Bayesian framework, the infection risk $r_i(t)$ is assumed to follow a posterior beta distribution $\pi_i(t) = Beta\big(1/\mathcal{B} + \sum_{\tau=t-k_0+1}^{t} n_i(\tau), (1 - 1/\mathcal{B}) +$

$\sum_{\tau=t-k_0+1}^{t}(q_i(\tau)-n_i(\tau))$), where $k_0 \in \mathbb{N}$ is the length of a sliding window incorporating all the latest k_0-day testing data for parameter update to ensure robust and effective assessment of infection risk.

Since the disease is more likely to transmit in the BG with higher infection risk, we develop a physics-informed model incorporating multiple factors related to two physical natures of transmission dynamics and health disparity embedded in contagion of infectious diseases to project prospective infection risk, modeled by $\boldsymbol{R}(t) = \boldsymbol{Z}(t)\boldsymbol{r}(t)$, where $\boldsymbol{Z}(t) = v_0\boldsymbol{T} + v_1\boldsymbol{K}(t)$ is a symmetric transmission matrix that accommodates two transmission patterns both within and between BGs with two scaling parameters v_0, v_1: (i) Through a power transformation function $T_g(\cdot)$, $\boldsymbol{T} = \text{diag}\left(T_g(l_i)\right)$ accommodates the local transmission pattern within each BG itself, evaluated by the Local Contagion Risk Score (LCRS) l_i; (ii) Through a kernel function $K_{h_i}(\cdot)$, $\boldsymbol{K}(t) = \left[K_{h_i}\left(\frac{1}{c_{i,j}(t)}\right)\right]_{|G|\times|G|}$ accommodates the interaction transmission pattern among different BGs evaluated by the dynamic BG Connectedness Score (CS) $c_{ij}(t), \forall i,j \in G$, as the mobility of citizens varies over time during the spread of the pandemic. On top of the physics-informed model, $\boldsymbol{R}(t)$ further accounts for spatial and temporal transmission patterns of infectious diseases, thus implying the prospective infection risk for each BG. Apart from the infection risk, the severity risk of the disease, which reflects vulnerability difference to a severe illness for each community due to the health disparity, is in practice also a crucial aspect for assessing necessity of tests allocation and should thus be rigorously considered, especially when the disease mortality rate is high. It can be similarly evaluated through a power transformation function of the Disease Severity Risk Score (DSRS) w_i by $H_f(w_i) = w_i^f$. The parameters in the physics-informed model will be estimated based on early testing data through an optimization problem, and an iterative gradient descent algorithm is developed to solve it. Medical geographic analysis is conducted to screen out the sociodemographic and socioeconomic factors that are highly related to COVID infection for score evaluation (LCRS, CS, and DSRS; refer to Appendix for details). All factors are readily available at the BG level throughout the nation [3, 4], ensuring that the proposed method can be directly scaled to most states and regions.

Next, we take both severity and infection risks into consideration to derive the BG risk level $M_i(t) = H_f(w_i)R_i(t)$, which indicates the risk level of getting severely infected for BG i at time t. High risk level implies high test allocation priority, and $\boldsymbol{M}(t)$ is thus adopted as an informative measure to determine

the allocation. Two powerful MAB approaches, TS and UCB algorithms, are adapted to the proposed method on top of the physics-informed model due to their superior exploitation-exploration balance. Two types of informativity statistic $V_i^{(\cdot)}(t)$ implying the BG risk level are derived respectively: (i) Based on the TS algorithm, draw a random sample $\boldsymbol{V}^{(TS)}(t) \sim N\left(\mathbb{E}[\boldsymbol{M}(t)], Var\left(\boldsymbol{M}(t)\right)\right)$; and (ii) Based on the Bayesian UCB algorithm, $\boldsymbol{V}^{(UCB)}(t) = \mathbb{E}[\boldsymbol{M}(t)] + e^{(UCB)}\sqrt{Var\left(\boldsymbol{M}(t)\right)}$. $\mathbb{E}[\boldsymbol{M}(t)]$ implying the risk levels facilitates exploitation on the vulnerable BGs while $Var\left(\boldsymbol{M}(t)\right)$ implying the uncertainty of risks facilitates exploration on the potentially suspected BGs, combining which two a high value of $V_i^{(\cdot)}(t)$ specifies allocation priority. As a result, both the TS and UCB-based allocation strategies enables the collection of high-quality data for effective surveillance across all the BGs. In addition, we note that in practice the allocation also depends on the BG population, especially when the population difference among BGs is significant. A variable size adaptive allocation strategy is thus designed by assigning Q proportional to both $\boldsymbol{V}^{(\cdot)}(t)$ and $\boldsymbol{B}$, such that updating $q_i^{(\cdot)}(t+1) = \min\left\{\frac{V_i^{(\cdot)}(t)B_i}{\sum_j V_j^{(\cdot)}(t)B_j}Q, B_i\right\}, \forall i \in G$.

Finally, based on the collected testing data, two real-time monitoring procedures are respectively developed using two local statistics $\boldsymbol{r}(t)$ and $\boldsymbol{R}(t)$ for quick detection of disease outbreaks across all the BGs, the development of which two procedures is similar. We first estimate the infection risk based on the latest testing data, and then identify the subset of suspected BGs with relatively high risks that may lead to a disease outbreak across all the BGs through hypothesis testing, denoted by $D^{(\cdot)}(t)$, where the superscript represents r- or R-based procedures. To accelerate outbreak detection, the monitoring statistic $S^{(\cdot)}(t)$ showing the overall risk across all the BGs at time t is developed based on the diagnosed subset $D^{(\cdot)}(t)$. Subject to limited access to well-labeled in-control (IC) historical data in practice, especially for a new infectious disease such as COVID-19, we regard the testing data before time $t - k_{IC}$ as pseudo-IC data and derive the one-sided control limit $K_t^{(\cdot)}$ accordingly through hypothesis testing. A disease outbreak across all the BGs will be declared when $S^{(\cdot)}(t) > K_t^{(\cdot)}$ for τ successive time points.

3 Simulation

To thoroughly evaluate the performance of the proposed method, 200-day simulation data are generated under three transmission cases based on the COVID-19 pandemic in North Central Florida subject to

different kinds of social distancing adherence during the spread of the infectious disease: (i) Case I: few individuals adhere to social distancing; (ii) Case II: the social distancing adherence increases over time; (iii) Case III: the social distancing adherence fluctuates over time. The overall positive rate across all the BGs over time under each case are visualized by the blue lines in Figure 1(a).

Based on the first 50-day simulation data, we first investigate the performance of parameter estimation of the physics-informed model through the proposed iterative algorithm, which is evaluated by the mean absolute errors (MAE) with the corresponding 95% confidence interval (CI). The results summarized in Table 1 with generally small MAEs indicate a satisfactory accuracy of the parameter estimation. Figure 1(a) further visualizes the specific progression of projected overall positive rates over time compared to the true ones, which not only shows that the calibrated physics-informed model based on estimated parameters is robust and reasonable but also demonstrates its practicality to accurately analyze and project transmission patterns of infectious diseases.

Table 1. Performance of the parameter estimation.

Case	**I**	**II**	**III**
MAE	0.0005 (0.0006)	0.0004 (0.0004)	0.0004 (0.0002)
95% CI	(0.0000, 0.0012)	(0.0000, 0.0007)	(0.0001, 0.0007)

On top of the calibrated physics-informed model, the allocation performance of the proposed TS and UCB algorithms is then evaluated and will be compared with three baseline algorithms: (i) the greedy algorithm focusing only on exploitation with $V_i^{(greedy)}(t) = \mathbb{E}[M_i(t)]$; (ii) the exploration (ER) algorithm focusing only on exploration with $V_i^{(ER)}(t) = 1$; and (iii) the random sampling (RS) method which randomly allocates tests to the entire population. Based on 100 simulation runs, Figure 1(b) shows the 95% CIs of the total positive cases over time by different algorithms respectively, from which we observe that the TS, UCB and greedy algorithms outperform the ER and RS methods for all cases in identifying positive cases. This is expected since both ER and RS methods focus on exploring all the BGs without sufficient tests allocated to the most suspected BGs. The greedy algorithm achieves the best performance with more positive cases in Cases I and II, which is not surprising as there is only a few suspected BGs at relatively high risks and this scenario matches with its focus on exploiting top BGs, which is yet impractical in reality. When there are many suspected BGs that may change over time like Case III, the greedy algorithm lacking exploration will be misleading and result in degraded allocation performance. In contrast, the proposed TS and UCB algorithms still achieve a robust and

satisfactory allocation performance due to the superior exploration-exploitation balance, thus increasing the quality of the testing data collected for quick outbreak detection.

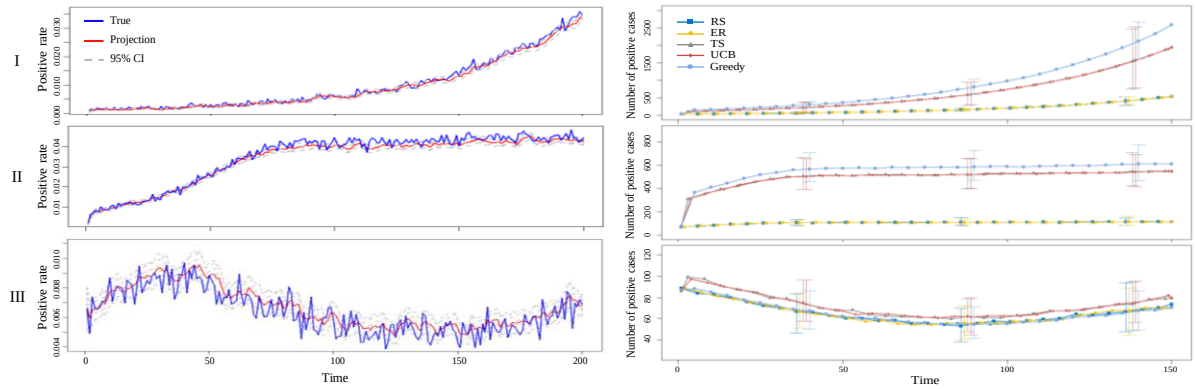

Figure 1. (a) Overall positive rates across all the BGs over time; (b) 95% CIs of total positive cases by different algorithms.

Finally, the performance of the real-time monitoring procedures in the proposed framework is also evaluated through comparison with the three baseline allocation algorithms in terms of four metrics. The simulation results of which tabulated in Table 2, and the findings are generally consistent with those in the allocation performance. This is expected since better allocation performance enables the collection of high-quality testing data that will thus lead to better monitoring performance for quick outbreak detection. As observed, although the monitoring performances of the greedy, TS and UCB algorithms are comparable on detection delays in Case II, in Case III the proposed UCB algorithm performs the best while the performance of the greedy algorithm deteriorates considerably. It can be also seen that the r-based procedures generally outperform the R-based ones especially in Case II with relatively high sensitivity and high specificity. This is probably because the R-based ones estimate risks by further making projections incorporating the transmission dynamics, such that high observed risk in a BG is spatially balanced out with neighboring BGs. This mechanism leads to a longer detection delay but will result in less false alarms in R-based procedures, which can be observed from Case III. Generally, the proposed TS and UCB algorithms integrated with the r -based monitoring procedures are overall preferred with reduced detection delay especially in cases of a small number of suspected BGs.

Table 2. Monitoring performance through different monitoring procedures.

			Sensitivity	Specificity	Detection delay	Number of false alarms
II	**TS**	**r**	**0.65 (0.05)**	**0.75 (0.06)**	**0.61 (0.49)**	**0**
		R	0.54 (0.07)	0.71 (0.06)	1.5 (1.57)	0
	UCB	**r**	**0.66 (0.05)**	**0.74 (0.06)**	**0.56 (0.50)**	**0**
		R	0.53 (0.07)	0.71 (0.07)	1.26 (1.59)	0
	Greedy	r	0.66 (0.05)	0.74 (0.06)	0.54 (0.50)	0
		R	0.04 (0.13)	0.98 (0.06)	133.73 (38.23)	0
	RS	r	0.56 (0.04)	0.55 (0.06)	15.93 (17.64)	0
		R	0.40 (0.04)	0.75 (0.06)	22.45 (30.71)	0
III	**TS**	r	0.98 (0.02)	0.18 (0.04)	-86.38 (9.86)	100
		R	**0.46 (0.07)**	**0.79 (0.04)**	**27.59 (21.66)**	**10**
	UCB	r	0.98 (0.02)	0.17 (0.04)	-85.21 (9.83)	100
		R	**0.51 (0.09)**	**0.76 (0.04)**	**19.58 (27.16)**	**15**
	Greedy	r	0.36 (0.06)	0.89 (0.03)	32.64 (21.47)	5
		R	0 (0)	1.00 (0.00)	50.14 (5.48)	0
	RS	r	0.99 (0.01)	0.14 (0.03)	-83.64 (8.03)	100
		R	0.36 (0.06)	0.89 (0.03)	38.45 (22.52)	7

4 Conclusions

In this work, we propose data-driven adaptive testing resource allocation strategies on top of a physics-informed model, which are integrated into a real-time monitoring framework which can effectively monitor infectious diseases in the context of limited testing availability. Despite inadequate testing data subject to limited tests, the proposed methodology is able to effectively assess health risks in different communities through the physics-informed model which accounts for different transmission patterns among communities by incorporating multiple factors related to transmission dynamics and health disparity that are innately embedded in contagion of infectious diseases. By further integrating the TS and UCB algorithms into the Bayesian framework, the proposed allocation strategies strike a superior exploration-exploitation balance between potentially suspected communities with high uncertainty of health risks and suspected communities at the highest risks, which adaptively and intelligently allocate the limited testing resources to collect high-quality testing data, thus enabling quick detection of disease outbreaks across all the communities as well as the diagnosis of most suspected communities that lead to the outbreak. A comprehensive simulation study is also conducted under three transmission cases to thoroughly evaluate the proposed methodology, from which the results show an overall robust and satisfactory performance of the proposed methodology in terms of the parameter estimation, the allocation strategies, and the real-time monitoring procedures.

Appendix: Score evaluation in the physics-informed model

In this appendix, we will demonstrate in detail how the three scores (LCRS, CS, and DSRS) in the physics-informed model in Section 2 are calculated, including what factors can be incorporated to evaluate the three scores to assess health risks in different communities. There are generally two steps for score evaluation: (i) collecting data and conducting medical geographic analysis to screen out the socioeconomic and sociodemographic factors that are highly related to the contagion of infectious diseases based on domain knowledge; (ii) providing and aggregating statistical descriptions of relevant factors for each type of scores followed by a prepossessing step to acquire final scores across all the communities used in the physics-informed model in the proposed method. It should be noted that such framework can be applied to different levels of community by incorporating factors at different levels.

In this appendix, we focus on demonstrating what BG level factors can be incorporated to evaluate the scores and types of socioeconomic data for statistical descriptions, which are summarized in Table 3.

In particular, the LCRS for each BG measuring the neighborhood contagion risk within each BG will be calculated using housing and population structure data from American Community Survey [4], given that the dynamics of local transmission are strongly shaped by local indoor crowding, especially housing crowding [5, 6]. Principal component analysis is adopted to incorporate eight potential factors which correlate with contagion risk at the community level: population density, median income, percentage of households with no car, percentage of single parents among households with children, percentage of population contributing to service labor (trade, retail, construction and transportation), and housing density (crowding). The CS measuring how frequently two BGs interact will be created using adjacency, commuting patterns between core and periphery, and mobility scores (as a proxy for stay-at-home and social distancing adherence). We use a novel social distancing micro-data metric—measuring the degree to which residents from individual neighborhoods have decreased their mobility, as a proxy for social distancing adherence [7]. This metric assumes a "home address" based on a cellphone night-time location, and includes several factors for each BG including the median number of minutes each cell phone user stayed at home. The DSRS measuring the probability of a severe disease within each BG will be calculated using data from the American Community Survey [4] and the Neighborhood Atlas [3]. Given that prior pandemics have disproportionally affected deprived communities and higher mortality is associated with lower socio-economic status [8], the DSRS will incorporate social determinants of health factors: socioeconomic status, area deprivation index, age structure, rurality, demographics, and prevalence of comorbidities (hypertension, diabetes, heart disease, kidney disease, obesity, advanced age, smoking, and chronic obstructive pulmonary disease).

Table 3. Score Evaluation

	Description	Relevant Factors	Data source
LCRS	Measuring the neighborhood contagion risk within each BG	E.g., population density, median income, percentage of households with no car, percentage of single parents among households with children, percentage of population contributing to service labor (trade, retail, construction and transportation), and housing density (crowding)	American Community Survey [4]
CS	Measuring how frequently two BGs interact	E.g., adjacency, commuting patterns between core and periphery, and stay-at-home and social distancing adherence (mobility scores as a proxy)	SafeGraph [7]
DSRS	Measuring the probability of a severe disease within each BG	Social determinants of health factors, e.g., socioeconomic status, area deprivation index, age structure, rurality, demographics, and prevalence of comorbidities	American Community Survey [4]; Neighborhood Atlas [3]

Throughout the simulation study in Section 3 specifically, the LCRS l_i is evaluated by both the population density and community infection risk (the deprivation and percentage of population that works at a food service or tourism service job). The CS c_{ij} is evaluated by the reciprocal of the point distance, and the stay-at-home rate within each BG is incorporated as a proxy of mobility [7] to dynamically measure $c_{ij}(t)$, given that the mobility patterns for each BG may vary over time due to quarantine measures that will affect the CS. The DSRS w_i is measured by normalizing Area Derivation Index [3] that reflects different severity risks among BGs caused by the health disparities. Due to highly insufficient data of severe cases subject to inadequate diagnostic testing data, we set $f = 1$ in $H_f(\cdot)$ to measure the severity risk in BG risk assessment for testing resource allocation and focus only on diagnostic testing data generation.

References

1. Unkel, S., Farrington, C. P., Garthwaite, P. H., Robertson, C., and Andrews, N. (2012). Statistical methods for the prospective detection of infectious disease outbreaks: a review. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *175*(1), 49-82.
2. Žilinskas, J., Lančinskas, A., and Guarracino, M. R. (2021). Pooled testing with replication as a mass testing strategy for the COVID-19 pandemics. *Scientific Reports*, 11(1), 1-7.
3. Kind, A. J., and Buckingham, W. R. (2018). Making neighborhood-disadvantage metrics accessible—the neighborhood atlas. *The New England journal of medicine*, 378(26), 2456.
4. U.S. Census Bureau. (2020). 2016-2020 American Community Survey 5-year Data Release. Retrieved from https://www.census.gov/newsroom/press-kits/2021/acs-5-year.html.
5. Rader, B., Nande, A., Adlam, B., Hill, A. L., Reiner, R. C., Pigott, D. M., ... and COVID-19 data working group. (2020). Crowding and the epidemic intensity of COVID-19 transmission. *MedRxiv*.
6. Qian, H., Miao, T., Liu, L., Zheng, X., Luo, D., and Li, Y. (2021). Indoor transmission of SARS-CoV-2. *Indoor Air*, *31*(3), 639-645.
7. SafeGraph, I.N.C. (2020). The source of truth for poi data and business listings.
8. Nayak, A., Islam, S. J., Mehta, A., Ko, Y. A., Patel, S. A., Goyal, A., ... and Quyyumi, A. A. (2020). Impact of social vulnerability on COVID-19 incidence and outcomes in the United States. *MedRxiv*.