

## Optimal allocation of limited test resources for the quantification of COVID-19 infections

Michail Chatzimanolakis<sup>a,\*</sup>, Pascal Weber<sup>a,\*</sup>, George Arampatzis<sup>a</sup>, Daniel Wälchli<sup>a</sup>, Ivica Kičić<sup>a</sup>, Petr Karnakov<sup>a</sup>, Costas Papadimitriou<sup>b</sup>, and Petros Koumoutsakos<sup>a</sup>

<sup>a</sup>Computational Science and Engineering Laboratory, ETH Zurich, Switzerland

<sup>b</sup>Department of Mechanical Engineering, University of Thessaly, Greece

\*Contributed equally

### Summary

The systematic identification of infected individuals is critical for the containment of the COVID-19 pandemic. Currently, the spread of the disease is mostly quantified by the reported numbers of infections, hospitalisations, recoveries and deaths; these quantities inform epidemiology models that provide forecasts for the spread of the epidemic and guide policy making. The veracity of these forecasts depends on the discrepancy between the numbers of reported, and unreported yet infectious, individuals. We combine Bayesian experimental design with an epidemiology model and propose a methodology for the optimal allocation of limited testing resources in space and time, which maximises the information gain for such unreported infections. The proposed approach is applicable at the onset and spread of the epidemic and can forewarn of a possible recurrence of the disease after relaxation of interventions. We examine its application in Switzerland; the open source software is, however, readily adaptable to countries around the world. We find that following the proposed methodology can lead to vastly less uncertain predictions for the spread of the disease, thus improving estimates of the effective reproduction number and the future number of unreported infections. This information can provide timely and systematic guidance for the effective identification of infectious individuals and for decision-making regarding lockdown measures and the distribution of vaccines.

**Keywords**— Bayesian optimal experimental design, epidemiology, COVID-19

### Introduction

The identification of unreported individuals infected by SARS-CoV-2 is critical for the quantification, forecasting and planning of interventions during the COVID-19 pandemic [1]. Currently the spread of the disease is mostly quantified by the reported numbers of infections, hospitalisations, recoveries and deaths [2]. These quantities inform epidemiology models that provide short term forecasts for the spread of the epidemic, help quantify the role of possible interventions and guide policy making. The veracity of these forecasts depends on the

discrepancy between the numbers of reported, and unreported yet infectious, individuals.

In recent months, the estimation of unreported infections has been the subject of several testing campaigns [3, 4]. Although there is valuable information being gathered, their estimates rely on testing individuals who are either already symptomatic or have been selected based on certain criteria (hospital visits, airport arrivals, geographic vicinity to researchers, etc.). Generic, randomised tests of the population are broadly applied, but they have been hampered either by delays [5] or by insufficient numbers of test kits [6]. There is broad recognition that efficient testing strategies are critical for the timely identification of infectious individuals and the optimal allocation of resources [7]. However, targeted testing entails bias and random tests require access to a high percentage of the population with commensurate high costs. The quality of the data, as well as the ways they are incorporated in the epidemiology models, is critical for their predictions and for estimating their uncertainties [8]. A way to minimise these uncertainties by suitably distributing in space and time a given number of test kits is the subject of this work. This optimal allocation of testing resources and the respective increase in the fidelity of forecasting models are essential to effective policy making throughout the pandemic.

Here, we present a methodology for the OPTimal Allocation of LImited Testing resourceS (OPALITS) that maximises the information gain over any prior knowledge regarding infections. The method relies on forecasts by epidemiological models with parameters adjusted through Bayesian inference as data become available through suitable surveys [9]. The forecasts are combined with Bayesian experimental design [10–12] to determine the optimal test allocation in space and time for various objectives (minimise prediction uncertainty, maximise information gain of unreported infections). We emphasise that the proposed OPALITS is applicable in all stages of the pandemic, regardless of the availability of data.

We employ the *SEI<sup>r</sup>I<sup>u</sup>R* model [13], which quantifies the spread of a disease in a country's population distributed in a number of communities that are interacting through mobility networks. The *SEI<sup>r</sup>I<sup>u</sup>R* model predicts the number of sus-

ceptible ( $S$ ), exposed ( $E$ ), infectious reported ( $I^r$ ), unreported ( $I^u$ ), and removed ( $R$ ) individuals from the population. Here we focus on Switzerland and consider its cantons as the respective communities. The model parameters are: the relative transmission rate between reported and unreported infectious individuals ( $\mu$ ), the virus latency period ( $Z$ ), the infectious period ( $D$ ) and the reporting rate ( $\alpha$ ). The transmission rate ( $\beta$ ) and the mobility factor ( $\theta$ ) are considered to be time dependent in order to account for government interventions. For all stages of the epidemic, the uncertainties of the model parameters are quantified and propagated using Bayesian inference. At the onset of the epidemic, the uncertainty was quantified through prior probability distributions. As data on daily infections become available, the uncertainty in model parameters was updated through Bayesian inference. The parameter probability distributions are used to propagate uncertainties in the model forecasts and can assist decision makers in quantifying risks associated with the progression of the disease. The proper quantification of uncertainty bounds in the model parameters has a profound effect on predictions of the disease dynamics [8]. Large uncertainty bounds around the most probable parameter values hinder the decision process for identifying effective interventions.

The OPALITS aims to assign limited test-kit resources to acquire data that would reduce the model prediction uncertainties. Minimising the uncertainty of the model parameters leads to more reliable predictions for quantities such as the reproduction number [14]. Moreover, the reduced model uncertainties help minimise risks associated with the decision-making process including timing, extent of interventions and probability of exceeding hospital capacity.

We quantify the information gain from these tests using a utility function [12, 15] based on the Kullback-Leibler divergence between the inferred posterior distribution and the current prior distribution of the model parameters. The prior can be formulated using the posterior distribution estimated from daily data of the infectious reported individuals up to the current date (see Materials and methods). Hence, at any stage of the epidemic, the OPALITS provides guidance on the time and location/community where testing needs to be carried out to maximise the expected information gain regarding infections in a population.

We demonstrate the simplicity and applicability of the present method in estimating the spread of the coronavirus disease in the cantons of Switzerland. We find that the OPALITS methodology outperforms non-specific, randomised testing of sub-populations throughout the COVID-19 pandemic. The proposed strategy is readily applicable to other countries and the employed open source software can readily accommodate different epidemiological models.

## Methods

The optimal time (day) and location (canton) for surveying a population to detect infectious individuals is determined via Bayesian optimal experimental design [12, 16, 17]. This optimal testing allocation (OPALITS) relies on combining Bayesian inference and utility theory with forecasting models of the epidemic. We remark that the OPALITS does not depend on a particular epidemiological model or type of data. The methodology is applicable at all stages of the epidemic (inception to re-occurrence). It can operate without data at the early stages of the pandemic and takes advantage of data available at later stages of the pandemic. The methodology is rendered computationally efficient using a sequential optimization algorithm [18].

tionally efficient using a sequential optimization algorithm [18].

**Bayesian Inference from randomized testing** We consider a testing campaign including a set ( $\mathbf{s}$ ) of surveys  $s_i = (k_i, t_i)$ ,  $i = 1, \dots, M_y$  performed in location  $k_i \in \mathcal{C}$  and on day  $t_i \in \mathcal{T}$ . These surveys measure a quantity of interest (QoI), that is denoted by  $\mathbf{y}(\mathbf{s}) = (y_1, \dots, y_{M_y})$ . Here,  $y_i$  is the number of unreported infectious individuals, measured through survey  $s_i$ . The QoI can be predicted by a model  $\mathbf{g}(\mathbf{s}, \boldsymbol{\vartheta}, \tilde{\boldsymbol{\vartheta}})$  (here the  $SEI^r I^u R$  epidemiological model) that depends on parameters of interest  $\boldsymbol{\vartheta} \in \mathbb{R}^N$  and nuisance parameters  $\tilde{\boldsymbol{\vartheta}} \in \mathbb{R}^{\tilde{N}}$ . The distinction between model and nuisance parameters is discussed in later sections. We note that both sets of parameters are uncertain and the proposed method aims to reduce the uncertainty only in the parameters of interest.

A stochastic error term  $\boldsymbol{\varepsilon}(\mathbf{s})$  links the model prediction with the QoI

$$\mathbf{y}(\mathbf{s}) = \mathbf{g}(\mathbf{s}, \boldsymbol{\vartheta}, \tilde{\boldsymbol{\vartheta}}) + \boldsymbol{\varepsilon}(\mathbf{s}). \quad (1)$$

The error  $\boldsymbol{\varepsilon}(\mathbf{s})$  is assumed to follow a zero-mean multivariate normal distribution  $\mathcal{N}(0, \Sigma)$  with covariance matrix  $\Sigma \in \mathbb{R}^{M_y \times M_y}$ . The elements of the covariance matrix ( $\Sigma_{s, s'}$ ) correspond to surveys taken at  $s = (k, t)$  and  $s' = (k', t')$  and are given by

$$\Sigma_{s, s'} = \sigma_t \sigma_{t'} \exp\left(-\frac{|t - t'|}{\tau}\right) \delta_{kk'}, \quad (2)$$

where  $\delta_{kk'}$  is the Kronecker delta, which is 1 for  $k = k'$  and 0 otherwise. The correlation time  $\tau \in [0.5, 3.5]$  is considered a nuisance parameter. These assumptions about the covariance imply that surveys in different locations are not correlated, while those in the same location have an exponentially decaying temporal correlation. The latter avoids clustering of surveys in small time intervals [19, 20]. The factor  $\sigma_t \in \mathbb{R}$  is assumed proportional to the expectation of the QoI, taken over all possible survey locations and over the range of model and nuisance parameters

$$\sigma_t = c \frac{1}{K} \sum_{i=1}^K \mathbb{E}_{\boldsymbol{\vartheta}, \tilde{\boldsymbol{\vartheta}}} [g(s_i, \boldsymbol{\vartheta}, \tilde{\boldsymbol{\vartheta}})], \quad (3)$$

where  $s_i = (i, t)$ . The parameter  $c \in [0, 0.25]$  is considered a model parameter. The expectation  $\mathbb{E}_{\boldsymbol{\vartheta}, \tilde{\boldsymbol{\vartheta}}}[\cdot]$  is taken with respect to all parameters  $\boldsymbol{\vartheta}$  and  $\tilde{\boldsymbol{\vartheta}}$  that follow the prior probability distribution with density  $p(\boldsymbol{\vartheta}, \tilde{\boldsymbol{\vartheta}}) = p(\boldsymbol{\vartheta})p(\tilde{\boldsymbol{\vartheta}})$ .

Under these assumptions, the conditional probability of  $\mathbf{y}$  on  $\boldsymbol{\vartheta}$ ,  $\tilde{\boldsymbol{\vartheta}}$  and  $\mathbf{s}$  is given by

$$p(\mathbf{y}|\boldsymbol{\vartheta}, \tilde{\boldsymbol{\vartheta}}, \mathbf{s}) = \frac{1}{\sqrt{(2\pi)^{M_y} |\Sigma|}} \exp\left(-\frac{1}{2} \mathbf{z}^\top \Sigma^{-1} \mathbf{z}\right), \quad (4)$$

where  $|\Sigma(\mathbf{s})|$  is the determinant of the covariance matrix  $\Sigma(\mathbf{s})$  and  $\mathbf{z} = \mathbf{y}(\mathbf{s}) - \mathbf{g}(\mathbf{s}, \boldsymbol{\vartheta}, \tilde{\boldsymbol{\vartheta}})$ .

In the present study, the QoI measured by a survey is the number of unreported infectious individuals in a particular canton on a particular date. This implicitly assumes that there are no restrictions on when the survey can be conducted and that there are no observational delays, which means the the QoI is instantaneously obtained. Both assumptions are not restrictive however. Restrictions on the possible survey dates can be accounted for by simply excluding those dates from the dates on which the utility function is evaluated. Also, a delay of one day (meaning that two days are needed to survey a canton  $k$ , starting from day  $t$ ) would mean that  $\mathbf{y} = (I_k^u(t) + I_k^u(t+1))/2$  is measured. In other words, when there is a delay the measured

quantity can still be mapped to a model quantity, which allows us to perform Bayesian inference. There are several types of measurements (Rapid testing [21], PCR [22], Schwabs [23]) being proposed for testing asymptomatic individuals. We emphasize that our methodology is compatible with any of these types. Data related issues such as uncertainties, test sensitivities and delays in processing can be accommodated in the Bayesian inference framework and in the input to the SEIR model.

**Expected Information Gain** The most informative surveys  $\mathbf{y}$  provide the least uncertainty in the estimates of the model parameters  $\boldsymbol{\vartheta}$ . Starting with a user-postulated prior distribution  $p(\boldsymbol{\vartheta})$ , Bayesian learning is used to update the uncertainties in the model parameters leading to a posterior distribution  $p(\boldsymbol{\vartheta}|\mathbf{y}, \tilde{\boldsymbol{\vartheta}}, \mathbf{s})$ , based on the information contained in the test data  $\mathbf{y}$ . The Kullback–Leibler (KL) divergence between the posterior  $p(\boldsymbol{\vartheta}|\mathbf{y}, \tilde{\boldsymbol{\vartheta}}, \mathbf{s})$  and the prior distributions  $p(\boldsymbol{\vartheta})$  of the model parameters measures the distance between the two distributions. Informative data produce posterior distributions that differ from the prior; greater differences lead to higher information gain. Therefore, the most informative data  $\mathbf{y}$  correspond to the testing strategy (measurement locations and times) with the highest information gain [15, 24].

The OPALITS is identified by maximizing a utility function [12, 25]. One choice is the KL divergence  $u(\mathbf{y}, \tilde{\boldsymbol{\vartheta}}, \mathbf{s}) = D_{\text{KL}}(p(\boldsymbol{\vartheta}|\mathbf{y}, \tilde{\boldsymbol{\vartheta}}, \mathbf{s})||p(\boldsymbol{\vartheta}))$  quantifying the information gain from the data [12]. However, since data are not available in the experimental design phase, the utility function is selected here to be the expected KL divergence  $\mathbb{E}_{\mathbf{y}|\tilde{\boldsymbol{\vartheta}}, \mathbf{s}}[u(\mathbf{y}, \tilde{\boldsymbol{\vartheta}}, \mathbf{s})]$  over all data generated by the model prediction error equation 1. Also, to account for the uncertainty in nuisance parameters  $\tilde{\boldsymbol{\vartheta}}$ , encoded in the prior distribution  $p(\tilde{\boldsymbol{\vartheta}})$ , the expectation is also taken with respect to  $\tilde{\boldsymbol{\vartheta}}$ , which results in the utility function [25]

$$U(\mathbf{s}) = \mathbb{E}_{\tilde{\boldsymbol{\vartheta}}} \left[ \mathbb{E}_{\mathbf{y}|\tilde{\boldsymbol{\vartheta}}, \mathbf{s}} [u(\mathbf{y}, \tilde{\boldsymbol{\vartheta}}, \mathbf{s})] \right] = \iiint \log \left( \frac{p(\boldsymbol{\vartheta}|\mathbf{y}, \tilde{\boldsymbol{\vartheta}}, \mathbf{s})}{p(\boldsymbol{\vartheta})} \right) p(\boldsymbol{\vartheta}|\mathbf{y}, \tilde{\boldsymbol{\vartheta}}, \mathbf{s}) d\boldsymbol{\vartheta} p(\mathbf{y}|\tilde{\boldsymbol{\vartheta}}, \mathbf{s}) d\mathbf{y} p(\tilde{\boldsymbol{\vartheta}}) d\tilde{\boldsymbol{\vartheta}}. \tag{5}$$

By using Bayes' theorem

$$p(\boldsymbol{\vartheta}|\mathbf{y}, \tilde{\boldsymbol{\vartheta}}, \mathbf{s}) = \frac{p(\mathbf{y}|\boldsymbol{\vartheta}, \tilde{\boldsymbol{\vartheta}}, \mathbf{s}) p(\boldsymbol{\vartheta})}{p(\mathbf{y}|\tilde{\boldsymbol{\vartheta}}, \mathbf{s})}, \tag{6}$$

the utility function can be simplified to

$$U(\mathbf{s}) = \iiint \log \left( \frac{p(\mathbf{y}|\boldsymbol{\vartheta}, \tilde{\boldsymbol{\vartheta}}, \mathbf{s})}{p(\mathbf{y}|\tilde{\boldsymbol{\vartheta}}, \mathbf{s})} \right) p(\mathbf{y}|\boldsymbol{\vartheta}, \tilde{\boldsymbol{\vartheta}}, \mathbf{s}) d\mathbf{y} p(\boldsymbol{\vartheta}) p(\tilde{\boldsymbol{\vartheta}}) d\boldsymbol{\vartheta} d\tilde{\boldsymbol{\vartheta}}. \tag{7}$$

Note that the expected utility only depends on the locations and times of the measurements via  $\mathbf{s}$ . The term  $p(\mathbf{y}|\tilde{\boldsymbol{\vartheta}}, \mathbf{s})$  is the model evidence given by

$$p(\mathbf{y}|\tilde{\boldsymbol{\vartheta}}, \mathbf{s}) = \int p(\mathbf{y}|\boldsymbol{\vartheta}, \tilde{\boldsymbol{\vartheta}}, \mathbf{s}) p(\boldsymbol{\vartheta}) d\boldsymbol{\vartheta}. \tag{8}$$

The choice of the prior distribution  $p(\boldsymbol{\vartheta})$  for the parameters allows to incorporate prior knowledge from epidemiology. If no information is available from data, a case encountered in the beginning of the infection, a uniform prior distribution can be assumed. Table S5 summarizes our choice of prior distributions for all the involved uncertain quantities. If data  $\mathbf{d}$  of the daily

number of reported infectious individuals is available, Bayesian inference can be used to inform the prior distribution, as described later on. In this case, the prior  $p(\boldsymbol{\vartheta})$  in equation 7 is replaced by the distribution  $p(\boldsymbol{\vartheta}|\mathbf{d})$  informed from the data  $\mathbf{d}$ .

In the present work, the assumed nuisance parameters are the correlation time  $\tau$  and the initial condition of the unreported infections in the cantons of Aargau, Bern, Basel-Landschaft, Basel-Stadt, Fribourg, Geneva, Grisons, St.Gallen, Ticino, Vaud, Valais and Zurich  $\mathbf{I}_{\text{IC}}^u = (I_{\text{AR}}^u, I_{\text{BE}}^u, I_{\text{BL}}^u, I_{\text{BS}}^u, I_{\text{FR}}^u, I_{\text{GE}}^u, I_{\text{GR}}^u, I_{\text{SG}}^u, I_{\text{TI}}^u, I_{\text{VD}}^u, I_{\text{VS}}^u, I_{\text{ZH}}^u)$ , with prior distributions  $\mathbf{I}_{\text{IC}}^u \sim \mathcal{U}([0, 50]^{12})$  and  $\tau \sim \mathcal{U}([0.5, 3.5])$ .

**Epidemiological Model** Here we employ the  $SEIR^uR$  epidemiological model [13] to forecast the dynamics of the coronavirus outbreak in Switzerland

$$\frac{dS_k}{dt} = -\frac{\beta S_k I_k^r}{N_k} - \frac{\mu \beta S_k I_k^u}{N_k} + \theta \sum_{l=1}^K \left( \frac{M_{kl} S_l}{N_l - I_l^r} - \frac{M_{lk} S_k}{N_k - I_k^r} \right) \tag{9}$$

$$\frac{dE_k}{dt} = \frac{\beta S_k I_k^r}{N_k} + \frac{\mu \beta S_k I_k^u}{N_k} - \frac{E_k}{Z} + \theta \sum_{l=1}^K \left( \frac{M_{kl} E_l}{N_l - I_l^r} - \frac{M_{lk} E_k}{N_k - I_k^r} \right) \tag{10}$$

$$\frac{dI_k^r}{dt} = \alpha \frac{E_k}{Z} - \frac{I_k^r}{D} \tag{11}$$

$$\frac{dI_k^u}{dt} = (1 - \alpha) \frac{E_k}{Z} - \frac{I_k^u}{D} + \theta \sum_{l=1}^K \left( \frac{M_{kl} I_l^u}{N_l - I_l^r} - \frac{M_{lk} I_k^u}{N_k - I_k^r} \right) \tag{12}$$

$$\frac{dN_k}{dt} = \theta \sum_{l=1}^K (M_{kl} - M_{lk}), \tag{13}$$

where  $S_k, E_k, I_k^r$  and  $I_k^u$  denote the number of individuals in canton  $k = \{1, \dots, K\}$  that are susceptible, exposed, reported infectious and unreported infectious, respectively. We denote by  $K$  the number of cantons (26 in Switzerland), by  $N_k$  the total population of the canton  $k$ , while the population mobility between cantons  $k$  and  $l$  is denoted by  $M_{kl}$  with values obtained from the Swiss Federal Statistical Office [26]. The model parameters are the transmission rate ( $\beta$ ), the relative transmission rate between reported and unreported infectious individuals ( $\mu$ ), the virus latency period ( $Z$ ), the infectious period ( $D$ ), the reporting rate ( $\alpha$ ) and the mobility factor ( $\theta$ ).

We employ different time-dependent expressions for the transmission rate and the mobility factor for each stage of the epidemic. Constants are chosen for the start of an epidemic while in the cases of monitoring of interventions, the following expressions are used:

$$\beta(t) = \begin{cases} b_0, & t \leq \delta_1 \\ b_1, & \delta_1 < t \end{cases}, \theta(t) = \begin{cases} \theta_0, & t \leq \delta_1 \\ \theta_1, & \delta_1 < t \end{cases}, \tag{14}$$

where  $b_0, b_1, \theta_0$  and  $\theta_1$  are the transmission rates and mobility factors before and after the intervention. Time  $t = 0$  corresponds to the 25<sup>th</sup> of February 2020, and  $\delta_1 = 21$  to the 17<sup>th</sup> of March 2020, when the lockdown was announced in Switzerland [27]. Finally, for the third case (monitoring of a second outbreak) we assume that

$$\beta(t) = \begin{cases} b_0, & 0 \leq t \leq \delta_1 \\ b_1, & \delta_1 < t \leq \delta_2 \\ b_2, & \delta_2 < t \leq \delta_3 \\ b_3(t), & \delta_3 < t \end{cases}, \theta(t) = \begin{cases} \theta_0, & 0 \leq t \leq \delta_1 \\ \theta_1, & \delta_1 < t \leq \delta_2 \\ \theta_2, & \delta_2 < t \leq \delta_3 \\ \theta_0, & \delta_3 < t \end{cases}. \tag{15}$$

As in equation 14,  $b_0$  is the transmission rate before the intervention while  $b_1 = c_1 b_0$  and  $b_2 = c_2 b_0$  with  $c_1, c_2 \in [0, 1]$  are the transmission rates after the two interventions. Similarly,  $\theta_0$  is the mobility factor before any interventions took place, while  $\theta_1 = c_3 \theta_0$  and  $\theta_2 = c_4 \theta_0$  with  $c_3, c_4 \in [0, 1]$  are the mobility factors after the two interventions. Moreover,  $\delta_1$  and  $\delta_2$  correspond to the days of the interventions. The day when the measures are loosened is denoted by  $\delta_3$ . After that day, the transmission rate is gradually increasing

$$b_3(t) = \min(b_2 + \lambda(t - \delta_3), b_0), \tag{16}$$

with  $\lambda \in [0, 0.03]$ , while the mobility factor regains its initial value of  $\theta_0$ .

**Estimation of the Expected Information Gain** The calculation of the expected utility from equation 7 is performed with Monte-Carlo integration. Samples from the prior distribution are denoted by  $\boldsymbol{\vartheta}^{(i)} \sim p(\boldsymbol{\vartheta})$  and by  $\tilde{\boldsymbol{\vartheta}}^{(i)} \sim p(\tilde{\boldsymbol{\vartheta}})$ , while samples on the measurement space are denoted by  $\mathbf{y}^{(i,j)} \sim p(\mathbf{y}|\boldsymbol{\vartheta}^{(i)}, \tilde{\boldsymbol{\vartheta}}^{(i)}, \mathbf{s})$ , where  $i \in \{1, \dots, N_\vartheta\}$  and  $j \in \{1, \dots, N_y\}$ . With these samples, an estimate of the expected utility is computed as

$$\hat{U}(\mathbf{s}) = \frac{1}{N_\vartheta N_y} \sum_{\substack{i=1 \\ j=1}}^{i=N_\vartheta \\ j=N_y} \left[ \log \left( \frac{p(\mathbf{y}^{(i,j)}|\boldsymbol{\vartheta}^{(i)}, \tilde{\boldsymbol{\vartheta}}^{(i)}, \mathbf{s})}{p(\mathbf{y}^{(i,j)}|\tilde{\boldsymbol{\vartheta}}^{(i)}, \mathbf{s})} \right) \right], \tag{17}$$

$$p(\mathbf{y}^{(i,j)}|\tilde{\boldsymbol{\vartheta}}^{(i)}, \mathbf{s}) := \frac{1}{N_\vartheta} \sum_{n=1}^{N_\vartheta} p(\mathbf{y}^{(i,j)}|\boldsymbol{\vartheta}^{(n)}, \tilde{\boldsymbol{\vartheta}}^{(i)}, \mathbf{s}),$$

In our implementation the samples  $\boldsymbol{\vartheta}^{(i)}$  and  $\tilde{\boldsymbol{\vartheta}}^{(i)}$ , ( $i = 1, \dots, N_\vartheta$ ), remain the same for different values of  $\mathbf{s}$ . Thus, the model evaluations  $\mathbf{g}(\mathbf{s}, \boldsymbol{\vartheta}^{(i)}, \tilde{\boldsymbol{\vartheta}}^{(i)})$  are only carried out once and are stored and used in the iteration process involved in the optimization. This allows to separate the computational cost of the model evaluation from the cost of computing the utility, which scales as  $\mathcal{O}(N_\vartheta^2 N_y)$ .

**Optimal Location and Time of Testing** We define the optimal survey times and locations as

$$\mathbf{s}^* = \arg \max_{s_1, \dots, s_{M_y}} \hat{U}(\mathbf{s}), \tag{18}$$

where  $\mathbf{s}^* = (s_1^*, \dots, s_{M_y}^*)$  with  $s_i^* = (k_i^*, t_i^*)$  denote the locations  $k_i^*$  and times  $t_i^*$  for the optimal surveys with  $i \in \{1, \dots, M_y\}$ . For a grid search, the associated computational cost is  $\mathcal{O}((KT)^{M_y})$  and thus grows exponentially with the number of surveys. This curse of dimensionality is avoided by using a sequential optimization method [18,28,29] to approximate the global optimum by iteratively solving

$$s_n^* = \arg \max_s \hat{U}_n(s), \quad \forall n = 1, \dots, M_y, \tag{19}$$

where  $s = (k, t)$  is the location and time to be estimated sequentially starting with  $n = 1$  and

$$\hat{U}_n(s) = \hat{U}(\mathbf{s}), \quad \mathbf{s} = (s_1^*, \dots, s_{n-1}^*, s). \tag{20}$$

Following this, we define the expected information gain for survey  $n$  as

$$\Delta \hat{U}_n(s) = \begin{cases} \hat{U}_1(s), & n = 1 \\ \hat{U}_n(s) - \hat{U}_{n-1}(s_{n-1}^*), & n > 1. \end{cases} \tag{21}$$

**Quantification of Uncertainty** A data informed prior  $p(\boldsymbol{\vartheta}|\mathbf{d})$  of the model parameters  $\boldsymbol{\vartheta}$  can be computed from available data  $\mathbf{d} = (d_1, \dots, d_{M_d})$ , collected at  $M_d$  locations and days. Here, available data  $\mathbf{d}$  refer to the daily number of reported infectious individuals and they are contrasted from the data  $\mathbf{y}$  of the number of unreported infectious individuals. The latter are obtained from testing strategies at selected populations using optimal experimental design. The data is mapped via a distinct model output  $f(s, \boldsymbol{\vartheta}, \tilde{\boldsymbol{\vartheta}})$  through the following error model

$$p(d_i|\boldsymbol{\vartheta}, \tilde{\boldsymbol{\vartheta}}, \nu) = \mathcal{NB}(d_i | f(s_i, \boldsymbol{\vartheta}, \tilde{\boldsymbol{\vartheta}}), \nu) \tag{22}$$

where  $\mathcal{NB}$  is the negative binomial distribution with mean  $f$  and dispersion  $\nu$ . Also,  $s_i = (k_i, t_i)$  is the location and time the data  $d_i$  for  $i = 1, \dots, M_d$  was collected. The choice of a different error model, compared to equation 1, is based on the assumption that the data are independent and identically distributed. Such an assumption would not be acceptable in the measurement model in equation 1, as it may result in uncorrelated measurements that can become clustered in small time intervals [19,20].

The data  $\mathbf{d} = (d_1, \dots, d_{M_d})$  are the daily number of reported infections per canton in Switzerland [30] which corresponds to the following model quantity

$$f(s_i, \boldsymbol{\vartheta}, \tilde{\boldsymbol{\vartheta}}) := \int_{t_i-0.5}^{t_i+0.5} \frac{\alpha}{Z} E_{k_i}(\tau) d\tau \approx \frac{\alpha}{Z} E_{k_i}(t_i). \tag{23}$$

The posterior distribution that will be used subsequently as a data informed prior is obtained using Bayes' theorem

$$p(\boldsymbol{\vartheta}, \tilde{\boldsymbol{\vartheta}}|\mathbf{d}) = \frac{p(\mathbf{d}|\boldsymbol{\vartheta}, \tilde{\boldsymbol{\vartheta}}) p(\boldsymbol{\vartheta}, \tilde{\boldsymbol{\vartheta}})}{p(\mathbf{d})}, \tag{24}$$

and is sampled with a nested sampling algorithm [14]. Note the difference to equation 6 and the optimal testing methodology, where we are interested to reduce the uncertainty in  $p(\boldsymbol{\vartheta}|\mathbf{y}, \tilde{\boldsymbol{\vartheta}}, \mathbf{s})$ , which excludes the nuisance parameters  $\tilde{\boldsymbol{\vartheta}}$ . For the dispersion parameter in equation 22, it is assumed that  $\nu = r f(s_i, \boldsymbol{\vartheta}, \tilde{\boldsymbol{\vartheta}})$ . The coefficient  $r$  is unknown and included in the parameter set, where  $r \sim \mathcal{U}([0, 2])$ .

The three inferences performed are summarized in table S5, which shows the involved model parameters in each case. The histograms for the found samples are shown in figures S1, S2, and S3.

We remark that, using the present methodology, the inferred date for the beginning of the intervention is  $\delta_1 = 22.5$ , which is the 18<sup>th</sup> of March 2020, corresponding well with the 17<sup>th</sup> of March 2020 on which the lockdown was introduced in Switzerland [27]. Moreover, we infer a significant reduction in the mobility factor, which indicates that traffic between cantons was also minimized. For the inference III we plot the fit using the inferred parameters in figure S4. The daily reported cases per canton are shown, together with the data used for the inference.

## Results

### Optimal allocation of limited testing resources (OPALITS) during the COVID-19 pandemic

We present the optimal test-kit allocation strategy for three stages of the epidemic: (i) starting phase (blue), (ii) containment after enforcement of interventions (red) and (iii) relaxing of interventions and monitoring for a possible second outbreak

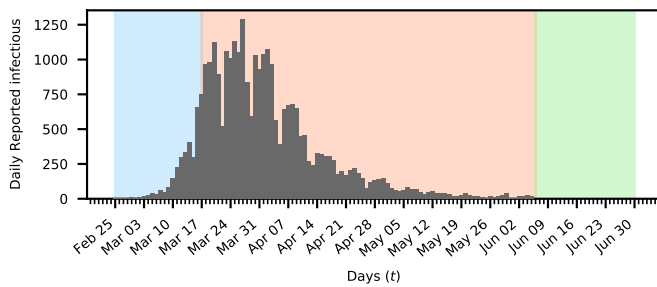


Figure 1: **Testing scenarios for the COVID19 outbreak in Switzerland.** Daily reported Coronavirus cases in Switzerland are plotted as gray bars. The period before (blue), during (red) and after (green) imposing non-pharmaceutical interventions are marked with color.

(green) (fig.1). The strategy relies on Bayesian experimental design and can operate when no data are available (as in the start of the epidemic), as well as when data have been accumulated, as in the last two stages of the epidemic. Testing campaigns rely on acquiring randomized samples from a population. The collected data, together with epidemiological models, help determine quantities of interest, such as the basic reproduction number of the disease [14]. By suitably adapting the testing campaign, the data can help to reduce the model uncertainty, thus enabling improved estimates regarding the severity of the epidemic.

A testing campaign consists of a set  $\mathbf{s}$  of surveys  $s_i = (k_i, t_i)$  which are labeled by  $i = 1, \dots, M_y$  and performed in locations  $k_i \in \mathcal{C}$  and on days  $t_i \in \mathcal{T}$ , where  $\mathcal{C}$  and  $\mathcal{T}$  are the set of all available locations and days, respectively. In this paper, a survey aims to determine the number of unreported infectious individuals in a particular location on a particular day. In the following we assume limited testing resources, where  $N$  test kits are available and each test kit corresponds to testing one person. The goal is to allocate these test kits in different times and locations so that we maximise the information gain regarding forecasts of the epidemiology model. The locations are the different Swiss cantons, and  $\mathcal{C} := \{\text{ZH, BE, LU, } \dots\}$  is the set of the strings with canton name abbreviations.

The results of the survey in a canton enable the estimation of a desired quantity of interest, such as the size of the unreported infected population ( $I^u$ ). The number of samples needed to estimate population proportions within a given confidence interval, error tolerance, and probability of proportion is given by Cochran's formula [9] corrected for a finite population size. Using Cochran's formula with confidence level 99%, error tolerance 1% and probability of infection 0.1 we find that the samples that would be required to survey the largest Swiss canton (Zurich) are approximately 5950. All the other cantons need up to 14% fewer samples, with the exception of the smallest canton that needs 27% fewer samples (supplementary fig. S7 in appendix 1). Hence we assume the minimum sample size is the same for all cantons. Assuming random sampling of a population with higher probability (up to 0.9) of infection or requiring tighter error bounds would have implied even more samples according to Cochran's formula. We note that as of October 2020, 1500 tests per one million people are performed on a daily basis in Switzerland [31]. This amounts to approximately 460 individual tests per canton, which is about an order of magnitude less than what would be required from Cochran's formula for informative random sampling. In turn, by using the proposed OPALITS, we can compensate for this

lack of test kits with an optimal and systematic process.

We outline the application of the proposed approach to a country with distinct administrative units (cantons in the case of Switzerland) (see figure 2). First, we determine how many cantons will be surveyed, given the number of available test-kits  $N$ . Then, the sequential optimisation of the expected utility function is performed (see Materials and methods) to identify optimal survey locations (cantons). We then distribute the test kits to the identified cantons and test a random subset of their population on the suggested day. After collecting the results from all the surveys we update the prior distributions of the model parameters. The collected data lead the maximal information gain in the model parameters. This in turn translates into minimal uncertainty in predictions made with the model for quantities, such as the number of unreported infections.

The expected information gain of a particular strategy for selecting the survey locations/times  $\mathbf{s}$  is quantified by a utility function  $\hat{U}(\mathbf{s})$  [15]. The maximum of this function corresponds to an optimal strategy that yields the most information about the quantities of interest. The expected utility function can be understood as a measure of the difference between prior knowledge of the model parameters and the posterior knowledge, after surveys have been conducted in a set of locations and dates. Given such a set, the utility function estimates the expected difference, the equivalent information gain, by taking the expectation over all possible survey results.

The OPALITS relies on forecasts by suitable epidemiological models. In turn, these forecasts rely on prior information and their predictions are further adjusted as data become available in a Bayesian inference framework [32]. The set of ordinary differential equations (ODEs) describing the  $SEI^u I^r R$  model [13] are integrated to produce the model output. The uncertainty of the model output and its discrepancy from the available data is quantified through a parametrised error model. The resulting stochastic model and its quantified uncertainties are then used to identify the optimal spatiotemporal allocation of limited test resources.

**Case 1: Beginning of the epidemic – optimal testing without data** At the start of an epidemic, there are no data and we assume no other prior information regarding the spread of the pathogen in a country. The initial conditions for the number of unreported infections ( $I_{IC}^u$ ) were selected with non-zero values for the cantons of Aargau, Bern, Basellandschaft, Basel-Stadt, Fribourg, Geneva, Grisons, St Gallen, Ticino, Vaud, Valais and Zurich based on their populations and their large number of interconnections. Because of the lack of any prior information and relevant data, all the parameters are assumed to follow uniform prior distributions (see table S5 in appendix 1 for details).

The first infectious person in Switzerland was reported on 25 February in the canton of Ticino ( $I_{T1}^u = 1$ ) with no initial reported infections in any other canton. The initial number of exposed individuals is set proportional to the number of unreported infections  $E_k = 3I_k^u$  in accordance with the value of  $R_0 \approx 3$  reported in [33] in the initial stage of the disease. The rest of the population is assumed to be susceptible. The methodology involves parameters of interest ( $\boldsymbol{\vartheta} = (\beta, \mu, \alpha, Z, D, \theta, c)$ ) and nuisance parameters ( $\tilde{\boldsymbol{\vartheta}} = (I_{IC}^u, \tau)$ ) that the testing strategy does not aim to determine (see Materials and methods section for definitions).

The estimated expected utility functions  $\hat{U}(\mathbf{s})$  for up to four surveys in the cantons of Switzerland for a time horizon of 8 days is shown in Figure 3,  $\mathcal{T} = \{\text{Feb 25}, \dots, \text{Mar 3}\}$ . Higher

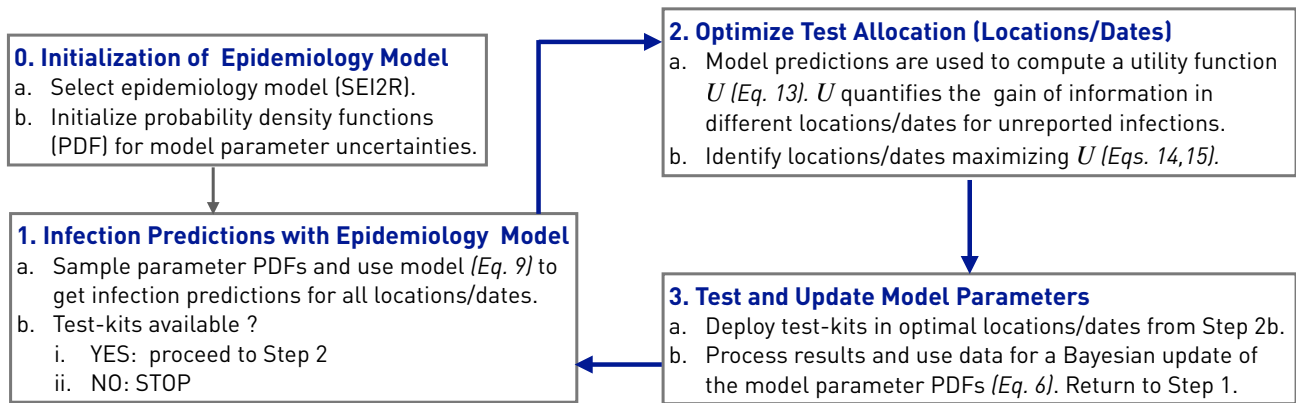


Figure 2: Schematic for the deployment of the Optimal Allocation of Limited Testing Resources (OPALITS) methodology

values for expected utility are estimated in cantons with larger population, reflecting the larger relative uncertainty for cantons with only few reported cases. This implies that smaller cantons, with lower mobility rates, are less preferred for performing tests since their contribution to the information gain is not significant. This reflects the fact that the assumed covariance matrix is shared among cantons (see Materials and methods). This implies a smaller relative error when surveying larger cantons with consequently higher number of infections. The Bayesian analysis allows the inference of the particular cantons and days on which a survey should be performed in order to maximise the information gain. Accordingly, the most informative survey should have been made in Zurich on 2 March. The optimal location and time for the second survey is determined to be canton of Vaud on the 27 February. As expected, the information gained from tests in the canton of Vaud is less than the information gained from the canton of Zurich. The information that would have been gained by surveying the next two selected cantons of Vaud and Basel-Landschaft on 3 March and 28 February, respectively, is progressively reduced to a small level that, given the testing costs, does not justify carrying out surveys in more than four cantons. The values of the optimal times are listed in table S1 in appendix 1.

The results indicate that the proposed OPALITS methodology selects certain populous and well interconnected cantons at specific times to acquire the most information for estimating the model parameters.

### Case 2: Exponential spreading and optimal testing strategy during nonpharmaceutical interventions

When the spreading of the coronavirus entered an exponential growth stage, several governments (including the Swiss) decided to make nonpharmaceutical interventions, such as requesting social distancing, closing schools and restaurants, or ordering a complete lockdown in order to contain the epidemic. Here, the goal of the OPALITS is to propose surveys that would help to better assess the effectiveness of these interventions.

In this case, probability distributions of model parameters are informed using data from the existing spread of the COVID-19. The daily reported infections in Switzerland [30] from the 25 February up to the 17 March 2020 are used to update the distributions specified in the previous phase by using Bayesian inference. The marginal posteriors are plotted in figure S1 (appendix 1). The  $SEI^r I^u R$  models the nonpharmaceutical interventions with a time-dependent transmission rate  $\beta$  and mobility factor  $\theta$ . These parameters are calibrated by

the data and provide an estimate of the timing and effectiveness of the interventions [8].

Figure 4 shows the maximum values of the information gain for each survey for  $\mathcal{T} = \{\text{Mar } 17, \dots, \text{Mar } 30\}$ . For cantons with a small population and low connectivity to other cantons, a low information gain is found. The opposite can be observed for cantons with large population and strong connections to other cantons. The values for the maximum utility in time for the measurements are listed in table S2 (appendix 1). If only a single canton were to be selected (due to limited availability of test kits in the country), then a survey in the canton of Vaud carried out on the 30 March would be preferred over surveys in the cantons of Zurich, Bern or Geneva (blue in fig. 4). If two surveys could be afforded, the OPALITS methodology proposes them in the same canton (Vaud) on the 17 and on the 30 March (blue and green in fig. 4). Note that the canton of Zurich, ranked as the next preferred canton for a single survey (blue in fig. 4), is not selected by the methodology since part of the information that would be gained from testing is already contained in surveys performed in Vaud. If more test kits were available, in addition to the two tests in Vaud, the optimal location and time for a third survey would have been the canton of Grisons on the 30 March (yellow in fig. 4). The canton of Zurich is proposed as the fourth location to be surveyed also on the 30 March. However, the information gain from the fourth survey in the canton of Zurich is approximately 10% of the total information gained from the surveys carried optimally in the first three cantons.

The results suggest that surveys at two locations/times provide significant information for assessing the effectiveness of interventions. Further tests on more locations/times did not add substantial information. It is evident that a trade-off between the required information gain and cost of testing are decisive for the number of necessary surveys and test kits.

**Case 3: Optimal monitoring for a second outbreak** After the relaxation of measures that assisted in mitigating the initial spread of the disease, it is critical to monitor the population for a possible second outbreak. The OPALITS methodology supports such monitoring with surveys of the population based on data up to and after the release of the measures.

First, Bayesian inference is performed with data available from 25 February to 6 June, to update the uniform priors the resulting marginal posteriors are shown in figure S2 (appendix 1). This date is in accordance with the first stage of the major release of measures in Switzerland [27]. The effects of interven-

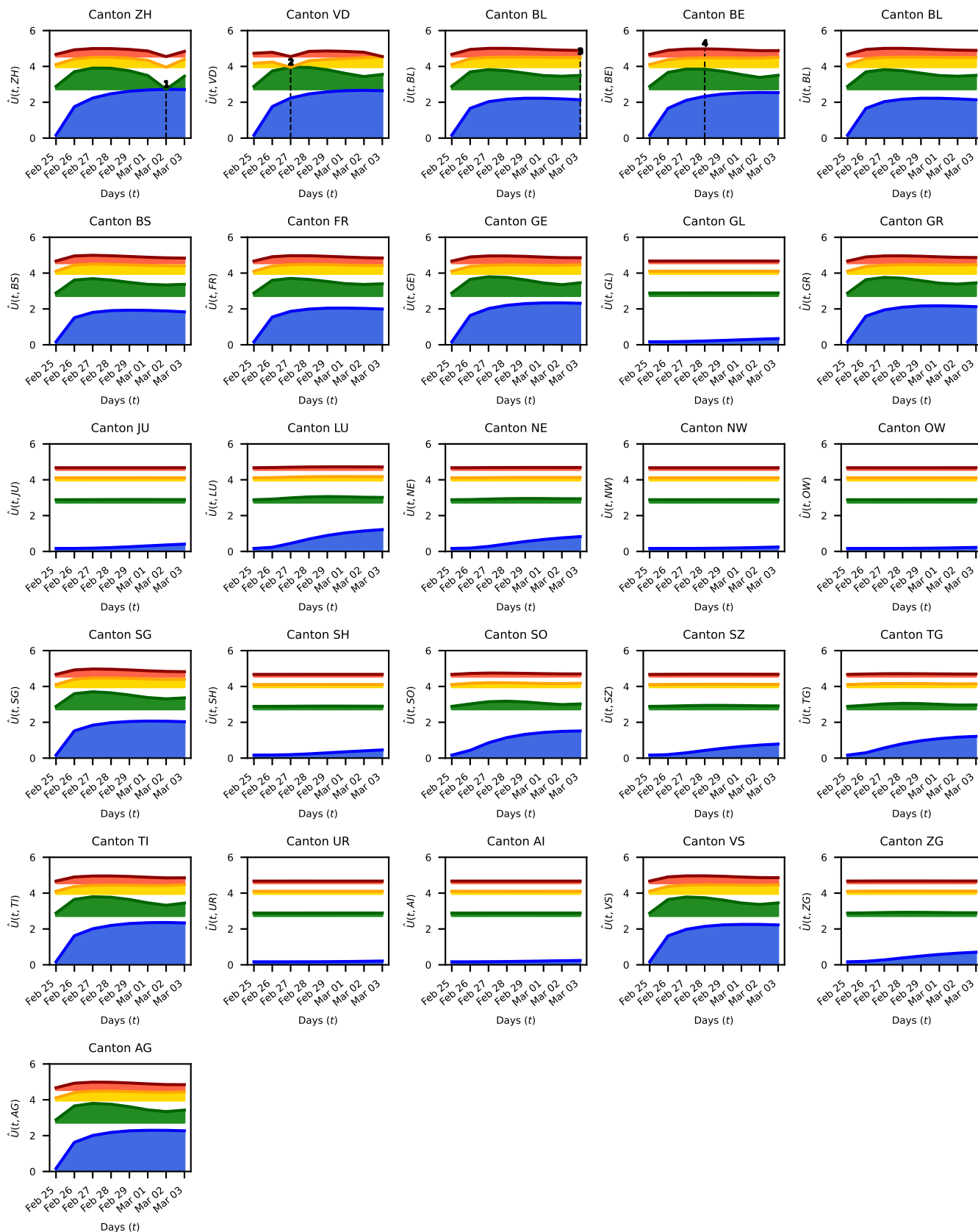


Figure 3: **Expected information gain during start of epidemic.** The blue curve corresponds to the utility of making one survey. The green curve is the utility when a second survey is added, provided that the location and time of the first survey correspond to the maximum of the blue curve (found in the canton of Zurich on 2 March). Similarly, the yellow and red curves show the utilities for a third and fourth survey, when the locations and times of the previous surveys are fixed to their optimal values. The fixed dates and location of each survey are plotted with black dashed lines. The shaded areas indicate the difference from the expected information gain of the previous survey, which becomes thinner as additional surveys do not yield a further significant information gain.

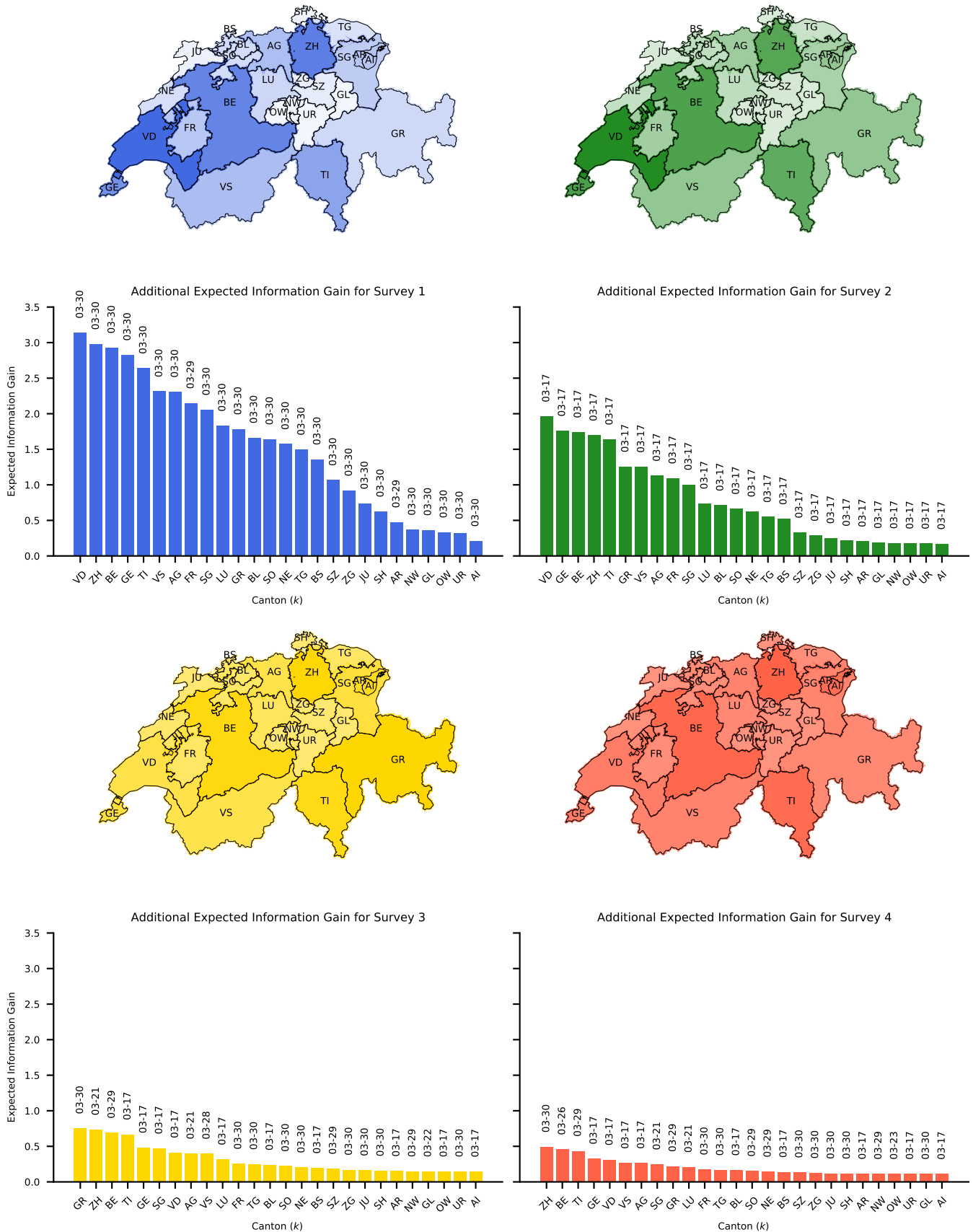
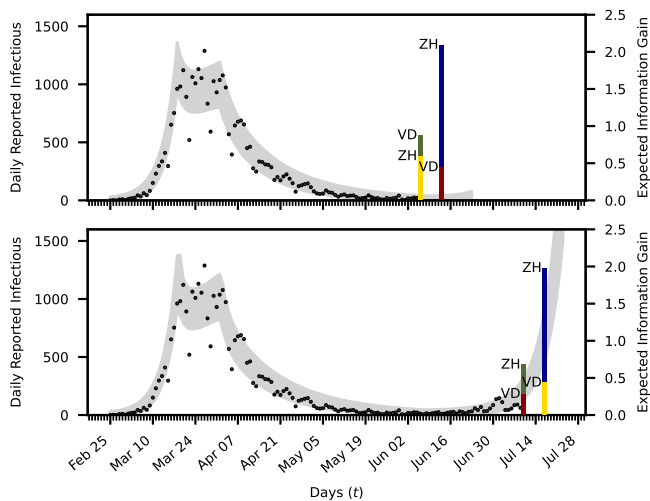


Figure 4: **Optimal testing strategy for effect of non-pharmaceutical interventions.** The maximum gain of information is plotted on the map of Switzerland using an exponential colour map. Here blue corresponds to taking one survey, green to adding a second, yellow to a third and red to a fourth. Below the map we plot the magnitude of the expected information gain of each survey, along with the optimal measurement dates per canton.





**Figure 5: Optimal testing strategy to monitor a second outbreak.** Bayesian inference determines the parameters of the first infection wave using the data (black dots) of the daily newly reported infections up to the 6 June (upper plot) and to the 9 July (lower plot). The 99% confidence intervals are plotted in grey. The proposed testing strategy is plotted with vertical bars at the optimal days found. Here blue indicated the utilities for the first survey. The green bars correspond to the gain in utility when adding a second survey assuming the first was chosen in the optimal location, where the yellow and red correspond to adding a third and fourth survey.

tions are modelled by a parametrised time-dependent transmission rate and mobility factor (see Materials and methods). The inferred probability distributions of these additional parameters are taken into account as the OPALITS maximises the information gain. Note that  $\mathcal{T} = \{\text{Jun } 7, \dots, \text{Jun } 14\}$  in this case.

Subsequently, data from 25 February to 9 July are included, repeating the Bayesian inference and estimating the marginal distributions and predictions shown in supplementary figures S3 and S4 (appendix 1),  $\mathcal{T} = \{\text{Jul } 10, \dots, \text{Jul } 17\}$ . The results indicate that the relaxation of measures correlates with an increase in the number of reported infections (fig. 5). The information gain for each canton indicates that the most informative surveys should be performed a week after performing the inference. The provided information could then assist in estimating the severity of a second outbreak, as indicated by the maximum of the utility in time (tables S3 and S4). Given that tests should be carried out in four locations and times, the methodology promotes optimal surveys for two different times, within a week, in the cantons of Zurich and Vaud. First, surveys should be performed in Zurich, providing high information gain for both considered cases. The next two surveys are to be performed in Zurich and Vaud, with a rank that depends on the considered case, and the fourth test should be performed in Vaud. We find that the information gain from the last test is approximately 10% of the cumulative information gain from the first three surveys. The number of surveys can be then selected according to the available test-kits  $N$ .

**Case 4: Effectiveness of optimal testing** We demonstrate the importance of following the OPALITS by comparing it with a non-specific testing campaign that is based on heuristics. We first re-examine the situation at the start of an

epidemic and assume that the available resources allow for two surveys. Surveys are simulated by evaluating the epidemiological model with the maximum a-posteriori estimate (MPE) of the parameters obtained from the inference in phase II (exponential growth) of the epidemic. We used data for the first 21 days of the infection spread in Switzerland [30] (25 February to 17 March). After evaluating the model, artificial surveys are obtained by adding a stochastic error term.

For the optimal strategy, data are collected by consulting figure 3. Thus, the two surveys are performed in the cantons of Zurich and Vaud, on the 2nd of 2 March and the 27th of 27 February, respectively. For a non-specific strategy, the cantons of Ticino and Bern were selected, on the 28th of 28 February. We remark that these are the canton where the first infection was reported and the capital of the country, respectively. These artificial data, obtained for the two strategies, are added to the real data of the daily reported cases from the first 8 days after the outbreak in Ticino. For the expanded data-set  $\mathcal{D}$  the posterior distributions  $p(\boldsymbol{\theta} | \tilde{\boldsymbol{\theta}}_{\text{MPE}}, \mathcal{D})$  are found by sampling the model parameters using nested sampling [14].

The resulting one- and two-dimensional marginalised posterior distributions for both strategies are shown in figure 6. We note that the dispersion coefficient  $r$  (defined in the Materials and methods) in the error model for the real data (the reported infections) and the correlation parameter are almost the same for both strategies. However, the model parameters show significant differences even when only two new data-points are added to a set of 208 data-points. The posterior distributions of the parameters of interest are propagated through the epidemiology model to provide the uncertainties in the number of unreported infectious individuals. In figure 7 the model output for the total number of unreported infections is plotted together with a 99% confidence interval along with the true value of the unreported cases obtained by using the selected parameters. The predictions from the OPALITS have a much higher certainty with a confidence interval that is up to four times narrower than the one from a non-specific strategy. The same figure also shows the relative histogram plots for the effective reproduction number, which for the employed model is given from  $R_t = \beta D \alpha + \beta D \mu (1 - \alpha)$  [13]. Not only is the histogram more peaked, when data are optimally collected, but also the mean value of the two histograms is different. When data are optimally collected, the found mean value for the effective reproduction number is 2.1, whereas when the non-specific strategy is followed the average value is 3.2. A mean value of 3.2 could lead to more strict non-pharmaceutical interventions, which might prove unnecessary and harmful for the economy.

Further comparisons, demonstrating the value of the OPALITS, include model predictions with higher certainty, as indicated by confidence intervals that are narrower than the ones obtained from a non-specific strategy (figs S5 and S6, appendix 1). Narrower uncertainty bounds provide higher confidence for decisions related to possible interventions to contain the epidemic.

## Discussion

We introduce a systematic approach to identify optimal times and locations for epidemiological surveys to quantify infectious individuals in a country's population during the COVID-19 epidemic. The proposed OPALITS methodology exploits prior information and available data to maximise the expected information gain in quantities of interest and to minimise uncertainties in the forecasts of epidemiological models.

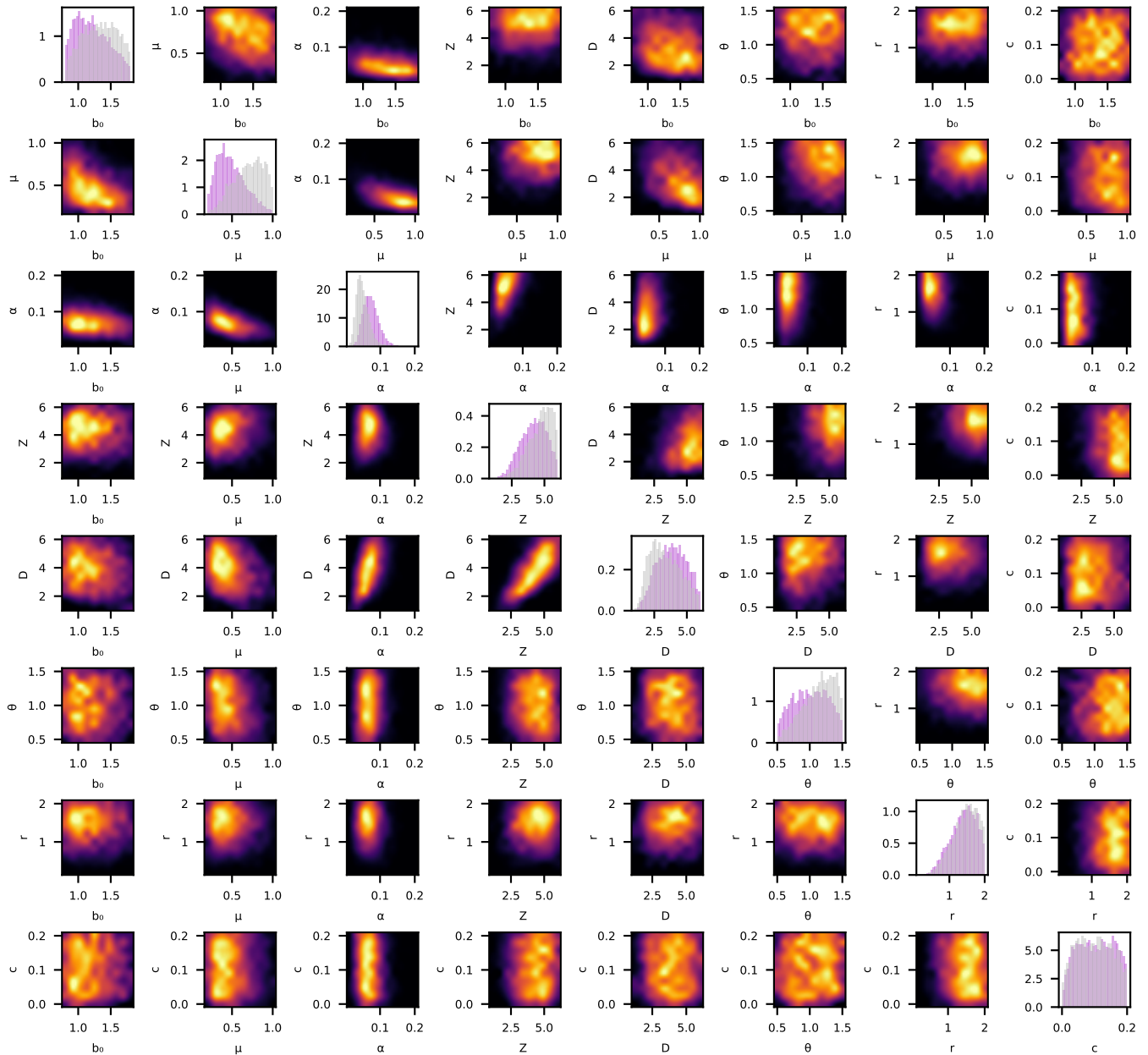
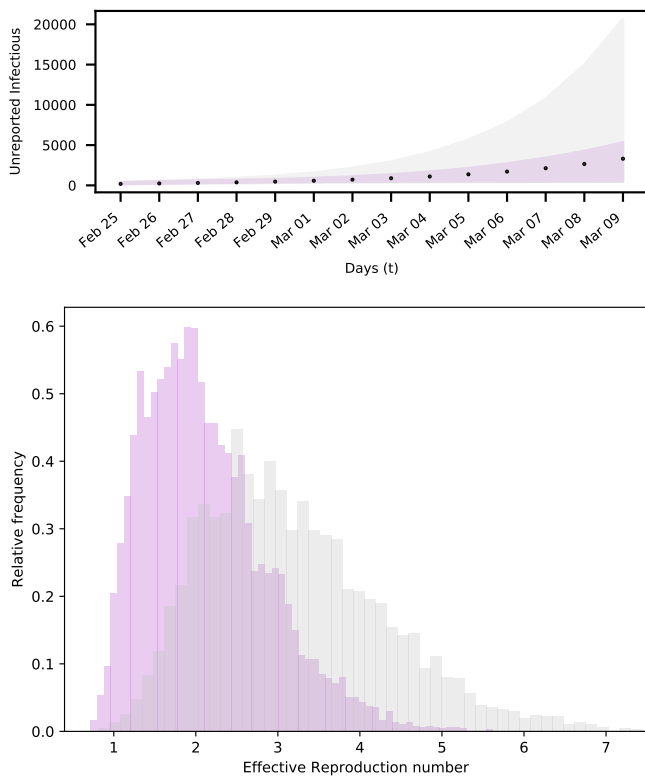


Figure 6: **Marginal posterior distributions for two strategies.** The diagonal shows the histogram for the marginal distribution for every parameter. Purple indicates posterior for the survey following the optimal testing strategy, gray the one for the non-specific strategy. The lower half and upper half show the samples of the joint distribution of two parameters for the optimal and the non-specific strategy respectively. Here black indicates low density and yellow high density.



**Figure 7: Prediction uncertainty for different testing strategies.** Up: The black dots show the actual unreported infectious for an artificial spread in Switzerland. The error bounds show the 99% confidence intervals of the model output for samples of the parameters with data obtained by optimal (purple) and non-specific testing (gray). Down: Relative frequency histograms for effective reproduction number, predicted with data obtained by optimal (purple) and non-specific testing (gray).

The present study addresses the need for an accurate assessment of COVID-19 infections [34] and it is shown to be far more accurate than the currently applied random testing. The proposed methodology is, to the best of our knowledge, the first method to propose an optimal spatiotemporal allocation of limited test kit resources. A first study of the estimation of unobserved COVID-19 infections [5] in the USA indicated that early testing would have decreased the surveillance gap during a critical phase of the epidemic. More recently, a number of studies have emerged that address the optimal allocation of resources. The “Test and Contain” process suggested in [7] addresses an idealised population of 10,000 and solves an allocation problem using predictions of the SIR model. They assume isolation of the positively identified individuals and showed that just one test a day can reduce the peak of infected individuals by 27%. This study is similar to ours in casting the test allocation problem in an optimisation framework, using linear programming in contrast to information maximisation that we propose. However, their approach is not data informed and does not address a realistic country scenario. Another study [35] focused on test kit allocation in the Philippines. They use a statistical approach and non-linear programming to determine the optimal percentage allocation of COVID-19 test kits among accredited testing centres in the Philippines, aiming for an equitable chance for all infected individuals to be tested. Their goal of optimal percentage allocation differs from ours, which is optimal space and time allocation of test kits.

The proposed method is demonstrated by focusing on the outbreak of the epidemic in Switzerland. We compare OPALITS with random testing and demonstrate its advantages in producing forecasts with far reduced uncertainties. We note that the existing testing capacity of 1500 tests per million people in Switzerland can be better allocated than the ongoing random testing. Moreover we show that the present methodology will be of particular importance to countries with testing capacity that is far lower than that of Switzerland [31].

The methodology relies on Bayesian experimental design using prior information and available data of reported infections along with forecasts from the  $SEI^*I^uR$  model. We compute the optimal testing strategy for three phases of the epidemic. At the onset of the epidemic the method identifies the most crucial dates and locations for randomised tests in the country’s population. The deployment of OPALITS at this phase would have allowed authorities to perform randomised testing in a period of high uncertainty, well in advance of the disease outbreak. Moreover, the presented approach is applicable to any newly arising epidemic and can be used to identify important surveying locations and a general protocol of action, whenever an unknown disease starts to spread. In the case of COVID-19, such course of action would limit early inaccurate estimates of metrics such as the virus mortality rate, estimated around 3% in early March 2020 by the World Health Organization [36] and currently believed to be lower than 1% [37].

During the period of nonpharmaceutical interventions, the proposed strategy would help quantify their effectiveness, assisting decision-making for further interventions or retraction of measures that may be harmful to the economy. In this study, available data for the daily reported infections prior to any interventions, combined with the proposed methodology, indicated that conducting two surveys after measures are imposed is sufficient. This can help to identify the new virus dynamics quickly and adjust interventions accordingly. Similarly, the OPALITS can assist monitoring for a recurrence of the disease after preventive measures have been relaxed and help guide further planning of interventions. Since massive testing for a

new disease might not be a possibility during its first outbreak and cheap individual tests might become available only later, applying the proposed methodology at this point provides a useful guideline on how to use the individual tests to conduct large-scale surveys. For instance, in Switzerland it was not before mid-April 2020 that rapid COVID-19 tests were released on the market [38]. Collecting data for the reported cases before that and using it to inform the proposed approach to find an OPALITS (after cheap individual tests become available) that will be applied during a possible lockdown would be the suggested course of action in this case.

There are a number of issues that the model should be able to accommodate in the future. These include accounting for virological test sensitivity, delays in the reporting of the test results and bias in the estimate of the unreported infected individuals (Cochran's formula). Further developments may include models that account for different transmission dynamics in cantons, and the classical Bayesian inference methods may be replaced with Hierarchical Bayesian Method to account for heterogeneous data.

We remark that the proposed OPALITS does not depend on a particular type of data/model or to the country of Switzerland. The open source code is modular, scalable and readily adaptable to different scenarios for the epidemic and countries around the world. We believe that the present work can be a valuable tool for decision makers to allocate resources efficiently for testing the population, providing a reliable quantification of the spread of the disease and designing effective interventions. Finally the accurate estimation of the spread of the disease can guide the timely distribution of vaccines.

## Disclosure statement

No financial support and no other potential conflict of interest relevant to this article were reported.

## Acknowledgments

We acknowledge discussions with Fabian Wermelinger, Lucas Amoudruz, Martin Boden (ETHZ). Sergio Martin (ETHZ) provided technical assistance with the software.

## References

- [1] Du Toit A. Outbreak of a novel coronavirus. *Nature Reviews Microbiology*. 2020;18(3):123–123.
- [2] Verity Rea. Estimates of the severity of coronavirus disease 2019: a model-based analysis. *The Lancet Infectious Diseases*. 2020;20.
- [3] Department of Health and Social Care, UK. Real-time Assessment of Community Transmission findings; 2020. Online, accessed 2020-10-18. <https://www.imperial.ac.uk/medicine/research-and-impact/groups/react-study/real-time-assessment-of-community-protect\penalty\z@-transmission-findings/>.
- [4] Lohse S, Pfuhl T, Berkó-Göttel B, Rissland J, Geißler T, Gärtner B, et al. Pooling of samples for testing for SARS-CoV-2 in asymptomatic people. *The Lancet Infectious Diseases*. 2020;20(11):1231 – 1232.
- [5] Perkins TA, Cavany SM, Moore SM, Oidtman RJ, Lerch A, Poterek M. Estimating unobserved SARS-CoV-2 infections in the United States. *Proceedings of the National Academy of Sciences*. 2020;117(36):22597–22602.
- [6] Abdalhamid B, Bilder CR, McCutchen EL, Hinrichs SH, Koepsell SA, Iwen PC. Assessment of Specimen Pooling to Conserve SARS CoV-2 Testing Resources. *American Journal of Clinical Pathology*. 2020 04;153(6):715–718.
- [7] Jonnerby J, Lazos P, Lock E, Marmolejo-Cossío F, Ramsey CB, Sridhar D. Test and Contain: A Resource-Optimal Testing Strategy for COVID-19. In: *AI for Social Good Workshop*; 2020. .
- [8] Karnakov P, Arampatzis G, Kičić I, Wermelinger F, Wälchli D, Papadimitriou C, et al. Data-driven inference of the reproduction number for COVID-19 before and after interventions for 51 European countries. *Swiss Medical Weekly*. 2020;150.
- [9] Cochran WG. *Sampling Techniques*. Wiley, New York; 1963.
- [10] Chaloner K, Verdinelli I. Bayesian Experimental Design: A Review. *Statistical Science*. 1995 08;10(2):273–304.
- [11] Huan X, Marzouk YM. Simulation-based optimal Bayesian experimental design for nonlinear systems. *Journal of Computational Physics*. 2013;232(1):288 – 317.
- [12] Ryan EG, Drovandi CC, McGree JM, Pettitt AN. A Review of Modern Computational Algorithms for Bayesian Optimal Design. *International Statistical Review*. 2016;84(1):128–154.
- [13] Li R, Pei S, Chen B, Song Y, Zhang T, Yang W, et al. Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2). *Science*. 2020;368(6490):489–493.
- [14] Speagle JS. Dynesty: a dynamic nested sampling package for estimating Bayesian posteriors and evidences. *Monthly Notices of the Royal Astronomical Society*. 2020 Feb;493(3):3132–3158.
- [15] Lindley DV. On a measure of the information provided by an experiment [Article]. *Annals of Mathematical Statistics*. 1956;27(4):986–1005.
- [16] Berry D. Bayesian clinical trials. *Nature reviews Drug discovery*. 2006 02;5:27–36.
- [17] Gongora AE, Xu B, Perry W, Okoye C, Riley P, Reyes KG, et al. A Bayesian experimental autonomous researcher for mechanical design. *Science Advances*. 2020;6(15).
- [18] Papadimitriou C. Optimal sensor placement methodology for parametric identification of structural systems. *Journal of Sound and Vibration*. 2004;278(4):923 – 947.
- [19] Papadimitriou C, Lombaert G. The effect of prediction error correlation on optimal sensor placement in structural dynamics. *Mechanical Systems and Signal Processing*. 2012;28:105 – 127. *Interdisciplinary and Integration Aspects in Structural Health Monitoring*.
- [20] Simoen E, Papadimitriou C, Lombaert G. On prediction error correlation in Bayesian model updating. *Journal of Sound and Vibration*. 2013;332(18):4136 – 4152.
- [21] Chau CH, Strobe JD, Figg WD. COVID-19 Clinical Diagnostics and Testing Technology. *Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy*. 2020;40(8):857–868.

- [22] Long C, Xu H, Shen Q, Zhang X, Fan B, Wang C, et al. Diagnosis of the Coronavirus disease (COVID-19): rRT-PCR or CT? *European Journal of Radiology*. 2020;126:108961.
- [23] Heller L, Mota CR, Greco DB. COVID-19 faecal-oral transmission: Are we asking the right questions? *Science of The Total Environment*. 2020;729:138919.
- [24] Karni E, Schmeidler D. Utility theory with uncertainty. In: *Handbook of Mathematical Economics*. vol. 4. Elsevier; 1991. p. 1763–1831.
- [25] Ryan KJ. Estimating Expected Information Gains for Experimental Designs With Application to the Random Fatigue-Limit Model. *Journal of Computational and Graphical Statistics*. 2003;12(3):585–603.
- [26] Bundesamtes für Statistik. Commuters for work purposes; 2020. Online. <https://www.bfs.admin.ch/bfsstatic/dam/assets/8507281/master>.
- [27] Bundesamtes für Gesundheit. Coronavirus: Informationen vom Bundesamtes für Gesundheit; 2020. Online. <https://www.bag.admin.ch/bag/de/home/krankheiten/ausbrueche-epidemien-pandemien/aktuelle-ausbrueche-epidemien/novel-cov.html>.
- [28] Verma S, Papadimitriou C, Lüthen N, Arampatzis G, Koumoutsakos P. Optimal sensor placement for artificial swimmers. *Journal of Fluid Mechanics*. 2020;884:A24.
- [29] Weber P, Arampatzis G, Novati G, Verma S, Papadimitriou C, Koumoutsakos P. Optimal Flow Sensing for Schooling Swimmers. *Biomimetics*. 2020 03;5:10.
- [30] Kanton Zürich, Statistisches Amt. SARS-CoV-2 Cases communicated by Swiss Cantons and Principality of Liechtenstein (FL); 2020. Online. [https://github.com/openZH/covid\\_19](https://github.com/openZH/covid_19).
- [31] Hasell J, Mathieu E, Beltekian D, Macdonald B, Giattino C, Ortiz-Ospina E, et al. A cross-country database of COVID-19 testing. *Sci Data*. 2020;7(345).
- [32] Dehning J, Zierenberg J, Spitzner FP, Wibral M, Neto JP, Wilczek M, et al. Inferring change points in the spread of COVID-19 reveals the effectiveness of interventions. *Science*. 2020;369(6500).
- [33] Liu Y, Gayle AA, Wilder-Smith A, Rocklöv J. The reproductive number of COVID-19 is higher compared to SARS coronavirus. *Journal of Travel Medicine*. 2020 02;27(2).
- [34] Angulo FJ, Finelli L, Swerdlow DL. Reopening Society and the Need for Real-Time Assessment of COVID-19 at the Community Level. *JAMA*. 2020 06;323(22):2247–2248.
- [35] Buhat CAH, Duero JCC, Felix EFO, Rabajante JF, Mamplata JB. Optimal Allocation of COVID-19 Test Kits Among Accredited Testing Centers in the Philippines. *medRxiv*. 2020;0.
- [36] World Health Organization. Q&A: Influenza and COVID-19 - similarities and differences; 2020. Online. <https://www.who.int/emergencies/diseases/novel-coronavirus-2019/question-and-answers-hub>.
- [37] Stringhini S, Wisniak A, Piumatti G, Azman AS, Lauer SA, Baysson H, et al. Seroprevalence of anti-SARS-CoV-2 IgG antibodies in Geneva, Switzerland (SEROCoV-POP): a population-based study. *The Lancet*. 2020;396(10247):313 – 319.
- [38] Launch of the first rapid test for COVID-19 on the Swiss market; 2020. Online. <https://www.startupticker.ch/en/news/april-2020/launch-of-the-first-rapid-test-for-covid-19\protect\penalty\z@-on-the-swiss-market>.