ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Applied Statistics
Series C

Appl. Statist. (2018)
67, *Part* 4, *pp.* 743–789

Optimal treatment allocations in space and time for on-line control of an emerging infectious disease

Eric B. Laber, Nick J. Meyer, Brian J. Reich and Krishna Pacifici

North Carolina State University, Raleigh, USA

Jaime A. Collazo

US Geological Survey North Carolina Cooperative Fish and Wildlife Research Unit, and North Carolina State University, Raleigh, USA

and John M. Drake

University of Georgia, Athens, USA

[*Read before* The Royal Statistical Society *on Wednesday, March 14th, 2018,* Professor R. Henderson *in the Chair*]

Summary. A key component in controlling the spread of an epidemic is deciding where, when and to whom to apply an intervention. We develop a framework for using data to inform these decisions in realtime. We formalize a treatment allocation strategy as a sequence of functions, one per treatment period, that map up-to-date information on the spread of an infectious disease to a subset of locations where treatment should be allocated. An optimal allocation strategy optimizes some cumulative outcome, e.g. the number of uninfected locations, the geographic footprint of the disease or the cost of the epidemic. Estimation of an optimal allocation strategy for an emerging infectious disease is challenging because spatial proximity induces interference between locations, the number of possible allocations is exponential in the number of locations, and because disease dynamics and intervention effectiveness are unknown at outbreak. We derive a Bayesian on-line estimator of the optimal allocation strategy that combines simulation–optimization with Thompson sampling. The estimator proposed performs favourably in simulation experiments. This work is motivated by and illustrated using data on the spread of white nose syndrome, which is a highly fatal infectious disease devastating bat populations in North America.

Keywords: Infectious disease control; Optimal allocation strategy; Thompson sampling; Treatment allocation; White nose syndrome

1. Introduction

Dynamical systems on networks and more general spatial domains have proved to be an effective modelling tool in many areas of science (Strogatz, 2001). Applications include ecological food webs (Williams and Martinez, 2000), electrical power grids, cellular and metabolic networks (Kohn, 1999), the World Wide Web (Broder *et al.*, 2000) and human mobility (Truscott and Ferguson, 2012). Interest in network anatomy and function underlies the greater movement towards research on complex systems. Recently, there has been increased interest in trying to understand dynamical systems on networks with the objective of administering control over

Address for correspondence: Eric B. Laber, Department of Statistics, North Carolina State University, 5216 SAS Hall, 2311 Stinson Drive, Raleigh, NC 27695-8203, USA.
E-mail: laber@stat.ncsu.edu

© 2018 Royal Statistical Society 0035–9254/18/67743

some process evolving over the network. An important example is the control of an epidemic evolving on a network of individuals. Fatal emerging diseases such as West Nile virus (Kilpatrick, 2011), white nose syndrome (WNS) (Maher *et al.*, 2012), foot-and-mouth disease (Tildesley *et al.*, 2006) and severe acute respiratory syndrome (Hufnagel *et al.*, 2004) represent a serious threat to ecological and environmental systems and to human health. The effect of bat loss due to WNS is projected to produce several billions of dollars of agricultural costs per year (Subcommittee on Fisheries, Wildlife, and Oceans, 2011). Understanding the dynamics of these epidemics and providing tools to control them efficiently and effectively are of paramount importance.

A key component in controlling the spread of an epidemic is deciding where, when and to whom to apply an intervention. A treatment allocation strategy formalizes this process as a sequence of functions, one per treatment period, that map up-to-date information on the epidemic to a subset of locations to receive treatment. An optimal treatment allocation strategy optimizes the expectation of some cumulative outcome, e.g. the cumulative number of infected individuals, the geographic footprint of the disease, the estimated total cost of the disease or a composite of several important outcomes. Estimation of an optimal treatment allocation strategy for an emerging epidemic presents several major challenges:

(a) scarcity of data—at the onset of the epidemic there is little information about disease dynamics and typically no information on the effectiveness of potential treatments;
(b) scalability—the number of possible allocations is exponential in the number of locations; for example, in the problem of WNS, there are more than 1100 locations leading to more treatment allocations than can possibly be enumerated by using existing computing resources;
(c) interference—dependence between locations violates the no interference between experimental units assumption (Sobel, 2006; Hudgens and Halloran, 2008);
(d) a long time horizon—an epidemic can persist for decades before eradication, and thus an optimal treatment allocation strategy must adapt to evolving logistical constraints, technologies and system dynamics.

We propose an on-line estimator of the optimal treatment allocation strategy designed to overcome these challenges. At each time point, the method proposed draws a model from the posterior distribution over system dynamics models and the estimated optimal allocation strategy is the maximizer over a prespecified class of strategies of the mean outcome under this model. The system dynamics model and estimated optimal allocation strategy are updated each time new data are collected to provide a continually evolving strategy. Furthermore, the class of potential allocation strategies is chosen to reduce computational complexity when scaling to large decision problems and to ensure that logistical or feasibility constraints are satisfied. We show that the estimator proposed can scale to problems with more than 1000 nodes, four covariates per node, 15 treatment periods and about $O(10^{150})$ possible allocations at each time period.

The methodology proposed is related to the idea of a dynamic treatment regime in personalized medicine. A dynamic treatment regime is a sequence of decision rules, one per treatment stage, that map up-to-date patient information to a recommended treatment (Murphy, 2003; Robins, 2004; Schulte *et al.*, 2014). Thus, like a treatment allocation strategy, a dynamic treatment regime is a sequence of functions that is used to dictate treatment decisions over time. Furthermore, one approach to estimating a dynamic treatment regime is to model the mean outcome as a function of each regime in a prespecified class and then to take the maximizer over this class as the estimated optimal regime (Robins *et al.*, 2008; Orellana *et al.*, 2010; Zhao *et al.*, 2012, 2014, 2015; Zhang *et al.*, 2012a, b, 2013; Zhao *et al.*, 2014, 2015; Kang *et al.*, 2014).

However, despite these similarities, the challenges that were mentioned previously prevent direct application of methodology for dynamic treatment regimes to the problem of spatiotemporal treatment allocation. Methods for dynamic treatment regimes assume that the data comprise a large number of independent and identically distributed trajectories observed over time. In contrast, in the allocation problem, we observe a single observation over the spatial domain at each time point; hence there is no independent replication. Furthermore, existing methods for dynamic treatment regimes are designed for settings with a small number of treatment options at each treatment stage, e.g. between two and five, whereas, in the spatial allocation problem, there are an astronomically large number of potential treatments. There has been some research on continuous treatments in dynamic treatment regimes (Rich, 2013; Rich *et al*., 2014; Laber and Zhao, 2015); however, these methods heavily rely on smoothness of an outcome regression model across treatment values which does not apply in the treatment allocation problem.

Both estimation of dynamic treatment regimes and estimation of an optimal treatment allocation fall under the umbrella of reinforcement learning problems (Bertsekas, 1996; Sutton and Barto, 1998; Powell, 2007; Sugiyama, 2015). Our proposed estimator is an approximate variant of Thompson sampling (Thompson, 1933) wherein allocations are chosen with probability that is proportional to the posterior probability that they are optimal. Thompson sampling has been studied in the reinforcement learning literature primarily in its application to bandit problems (Scott, 2010; Chapelle and Li, 2011; Agrawal and Goyal, 2011, 2012, 2013; Kaufmann *et al*., 2012; Korda *et al*., 2013; Gopalan *et al*., 2014; Russo and Van Roy, 2014). Osband *et al*. (2013) and Gopalan and Mannor (2015) applied Thompson sampling to sequential decision problems modelled as Markov decision processes. However, these estimators require

(a) a finite set of system states and
(b) that a fixed allocation strategy be applied without adjustment for potentially long periods of time.

In the settings that we consider, the system state is continuous and high dimensional (making discretization impractical) and the application of a fixed suboptimal allocation strategy for a prolonged period is neither ethical nor feasible. For a comprehensive survey of Bayesian reinforcement learning see Ghavamzadeh *et al*. (2015).

The work proposed adds to the large literature on mathematical spatial transmission models for disease modelling and control. Within this literature, a common approach is to postulate a mathematical model of disease spread and then to use simulation experiments to evaluate and compare candidate intervention strategies (see Anderson *et al*. (1992), Hufnagel *et al*. (2004), Riley (2007), Hollingsworth (2009), Ma *et al*. (2009), Keeling and Rohani (2011) and references therein). These models have generated new insights into disease transmission and control strategies across a wide range of application domains including avian influenza (Le Menach *et al*., 2006; Jung *et al*., 2009), Chagas disease (Barbu *et al*., 2009, 2011), Ebola virus disease (Lekone and Finkenstädt, 2006; Kramer *et al*., 2016; Li *et al*., 2017); foot-and-mouth disease (Ferguson *et al*., 2001a, b; Keeling, 2005; Tildesley *et al*., 2006), human immunodeficiency virus and acquired immune deficiency syndrome (Jacquez *et al*., 1988; Korenromp *et al*., 2000), severe acute respiratory syndrome (Huang *et al*., 2004; Bauch *et al*., 2005; Hollingsworth *et al*., 2006) and smallpox (Kaplan *et al*., 2002; Bozzette *et al*., 2003; Ferguson *et al*., 2003; Kretzschmar *et al*., 2004), among others. The proposed Thompson sampling estimator relies on a working model for the underlying disease dynamics and thereby benefits from this rich literature on disease modelling. However, the estimator proposed is distinct from these approaches in that it

(a) considers on-line estimation of an optimal allocation strategy which requires balancing

the choice of treatment allocations that lead to maximal model improvement with the choice of treatment allocations that are optimal under the current estimated model,
(b) optimizes over a large (possibly infinite) class of allocation strategies and
(c) accommodates evolving resource and logistical constraints.

Thus, where much of the focus of mathematical disease modelling has been on building and validating high quality transmission models, our focus is on how to incorporate these models in optimal on-line treatment allocation.

In Section 2, we discuss one of the motivating problems for the work proposed: controlling the spread of WNS in bats. In Section 3, we formally define an optimal treatment allocation strategy by using potential outcomes and discuss the problem of interference. In Section 4, we develop our estimator of the optimal treatment allocation strategy and construct a class of strategies that are flexible but highly scalable. In Section 5, we evaluate the performance of the method by using simulation experiments. In Section 6, we apply the methodology to data on the spread of WNS. Future work and open problems are discussed in Section 7.

The data that are analysed in the paper and the programs that were used to analyse them can be obtained from

`http://wileyonlinelibrary.com/journal/rss-datasets`

2. White nose syndrome in bats

WNS is a disease that is caused by the fungus *Pseudogymnoascus destructans* (formerly *Geomyces destructans*) and predominately affects hibernating bats in North America (Blehert *et al*., 2009). An infected bat will present with a white fungus on its muzzle, ears and/or wings, and erratic behaviour during hibernation. The erratic behaviour during hibernation depletes fat reserves and expends valuable energy, resulting in low survival and death (Blehert *et al*., 2009). Mortality rates exceed 90% in some areas and more than 5.7 million bats have died because of WNS (Blehert *et al*., 2009; US Fish and Wildlife Service, 2015).

WNS was first recorded in Schoharie County, New York State, in 2006 (Blehert *et al*., 2009) and is now found in 25 states, five Canadian provinces, as far south as the Mississippi, and as far west as Missouri; Fig. 1. More than half of the 47 species of bats in the USA hibernate, making them vulnerable to exposure. Currently, two endangered species, the Gray bat, *Myotis grisescens*, and the Indiana bat, *Myotis sodalis*, as well as one threatened species, the Northern Long-eared bat, *Myotis septentrionalis*, are infected with WNS (Blehert *et al*., 2009). The ecological damage due to loss of bats and the speed of spread are unprecedented and the long-term damage is still considered to be immeasurable (Blehert *et al*., 2009). Short-term estimates of economic damage hover around \$3.7 billion year^{-1} mainly due to agricultural loss (Boyles *et al*., 2011). The estimated value of bats to the entire agricultural industry is \$22.9 billion year^{-1} not including many secondary effects and impacts, e.g. downstream effects of increased use of pesticides, predation effects on evolved resistance of insects to pesticides and genetically modified crops (Boyles *et al*., 2011).

Because of the economic and ecological effects, there is a tremendous need for a comprehensive national plan to control the spread of WNS. In 2009, the US Fish and Wildlife Service, along with state agencies and universities, convened to create a national plan for the control of WNS (US Fish and Wildlife Service, 2015). This plan outlines necessary actions for co-ordination and provides an overall template to prevent further spread of WNS. Although the national plan puts forth the first steps in co-ordination between states and other agencies, it does not explicitly provide a treatment plan or strategy to control WNS (Szymanski *et al*., 2009). Each state is left to

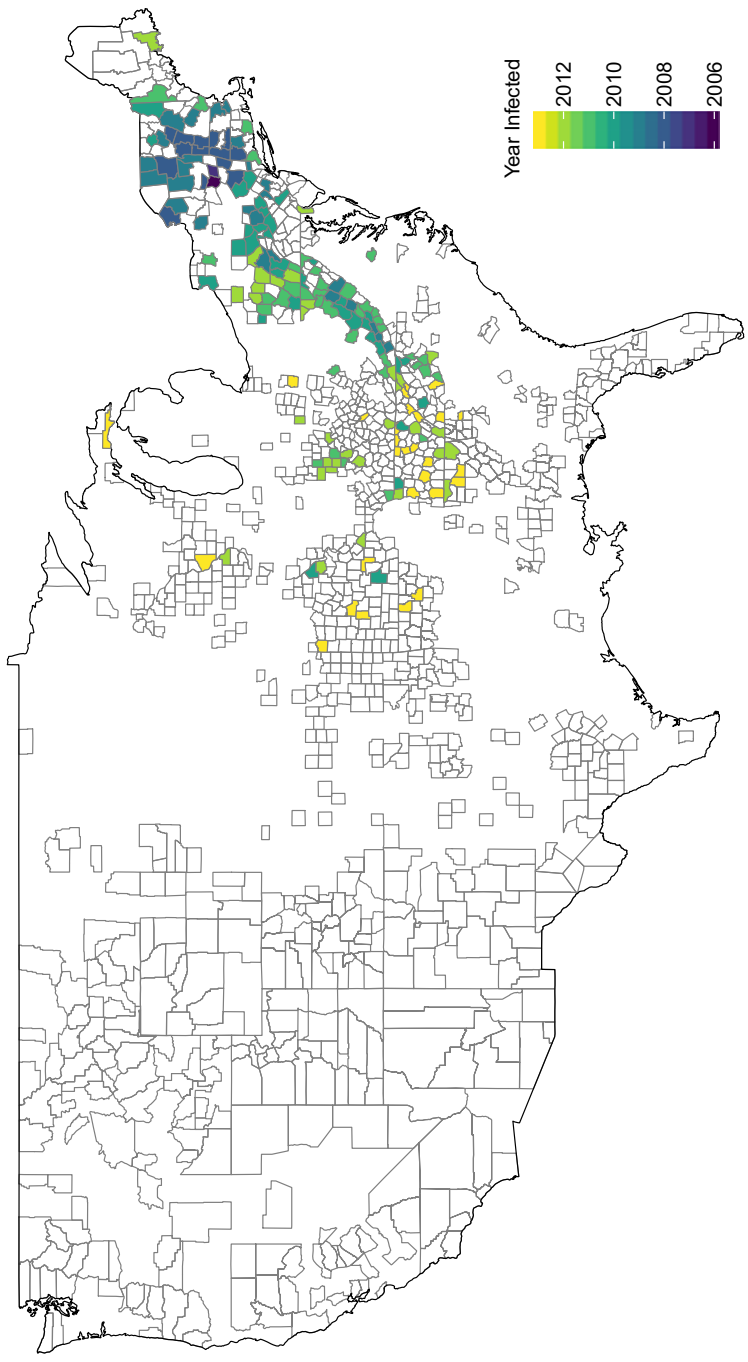

Fig. 1. Spread of WNS (US Fish and Wildlife Service, 2015): outlined counties contain caves; those without colour were uninfected at June 2014

implement treatments at its own discretion. Potential treatments include antifungal biological or non-chemical agents for bats at risk, modifying cave environmental variables, e.g. temperature and humidity, to slow fungus growth and to improve bat survival, vaccines to boost resistance and artificial caves (Cornelison *et al*., 2014; Hoyt *et al*., 2015). Unfortunately, many of these have not been tested in the field and their efficacy is currently unknown. Additional challenges exist because the disease has a highly complex nature of spread including a large spatial range (Maher *et al*., 2012). Therefore, to maximize what benefits these treatments may provide, it is essential to develop a principled, adaptive and data-driven control strategy that addresses the full potential range of WNS before further devastation occurs. We estimate such a control strategy and demonstrate that, if implemented, it may have a profound effect on the course of the current epidemic.

3. Defining an optimal treatment allocation strategy

We consider a decision problem evolving over a countably infinite set of treatment periods and a finite number of locations. The locations may represent physical locations in space, e.g. parcels of land identified as candidates for an intervention, or the locations may be nodes in a network, e.g. individuals in a social network. In the application to WNS, the data are provided at the county level, and thus cave bearing counties compose the locations of interest. At each time point, a decision maker observes information describing the current state of each location and subsequently uses this information to decide which locations should receive treatment. In the control of epidemics, location information would include information on the spread of the disease, e.g. infection status and time since infection among the infected, as well as features that are related to susceptibility or contagiousness. In WNS, for each county we observe the infection status, time since infected, number of caves, average winter temperature and a measure of species richness. For simplicity, our development considers the setting in which there are two possible choices at each location: apply a treatment or do nothing. However, the methodology proposed can be extended to handle settings in which several treatment options are available at each location. A treatment allocation strategy formalizes the treatment allocation process as a map from current information on all locations to a probability distribution over possible allocations. An allocation strategy is said to be optimal if it maximizes the mean cumulative utility over a prespecified class of strategies (minimizing cost can be handled in the obvious way).

Let $\mathcal{L} = \{1, \ldots, L\}$ denote the set of locations and $\mathcal{T} = \{1, 2, \ldots\}$ the set of treatment stages. The treatment stages may be dictated by the evolving decision process. Define $\mathbf{S}_l^t \in \mathbb{R}^p$ to be a summary of the information that is collected at location $l \in \mathcal{L}$ up to and including time $t \in \mathcal{T}$ and let $\mathbf{S}^t$ be $\{\mathbf{S}_l^t\}_{l \in \mathcal{L}}$; we assume that $\mathbf{S}^t$ is completely observed and measured without error. Let $A_l^t \in \{0, 1\}$ denote an indicator that location l received treatment at time t and $\mathbf{A}^t = \{A_l^t\}_{l \in \mathcal{L}}$ is the allocation at time t. Let $\mathcal{B}_L$ denote the set of all probability distributions over $\{0, 1\}^L$; and for a random variable U write $\text{supp}(U)$ to denote the support of U. A treatment allocation strategy π is a function from $\mathcal{S} = \text{supp}(\mathbf{S}^t)$ into $\mathcal{B}_L$ so that, under π, a decision maker who is presented with $\mathbf{S}^t = \mathbf{s}^t$ will select allocation $\mathbf{a}^t$ with probability $\pi(\mathbf{a}^t; \mathbf{s}^t)$. Allocation strategies of this type are termed stochastic strategies to contrast them with deterministic strategies which map states to allocations rather than to a distribution over allocations (Sutton and Barto, 1998). In the context of on-line estimation and optimization, the use of stochastic allocation strategies is critical to ensure consistent estimation of an optimal strategy (Kaelbling *et al*., 1996; Cesa-Bianchi and Lugosi, 2006) much in the same way that randomization is critical in adaptive clinical trials to ensure consistent estimation of an optimal treatment (Berry and Fristedt, 1985); see the on-line supplemental materials for an illustrative example. Let $Y_l^t \in \mathbb{R}$ denote an outcome

measured at location l at time t and let $\mathbf{Y}^t = \{Y_l^t\}_{l\in\mathcal{L}}$. For a prespecified constant $\gamma \in (0,1)$, the goal is to choose an allocation strategy that maximizes the mean of the discounted total utility $\Sigma_{t\geqslant 1}\gamma^{t-1}u(\mathbf{Y}^t)$, where $u(\cdot)$ is a scalar utility function and the constant γ balances proximal and distal outcomes. In some settings, it may be desirable to choose an alternative measure of cumulative utility, e.g. $\lim_{T\to\infty} T^{-1}\Sigma_{t=1}^{T} u(\mathbf{Y}^t)$; our methodology can be directly extended to handle such alternatives. We formalize the notion of an optimality allocation strategy by using potential outcomes (Rubin, 1978; Neyman, 1990).

Let Π denote a class of allocation strategies of interest; throughout, we implicitly assume that all allocation strategies under consideration belong to Π. Hence, the definition of optimality depends on Π. This class can be used to enforce logistical constraints, e.g. a limit on the number of locations that can be treated at each time point. Because our estimation algorithm is on line, this class of allocation strategies can be changed in realtime to reflect changing constraints. Define $\mathcal{F} = \{\mathbf{a} \in \{0,1\}^L : \mathbf{a} \in \text{supp}(\pi) \text{ for some } \pi \in \Pi\}$ to be the set of feasible allocations. We use overline notation to denote past history, e.g. $\bar{\mathbf{a}}^t = \{\mathbf{a}^v\}_{v=1}^t$, and an asterisk superscript to denote potential outcomes. For example, $Y^{*t}(\bar{a}^t)$ denotes the outcome that would be observed under treatment sequence $\bar{a}^t$. Define $\mathbf{W}^* = \{\mathbf{Y}^{*t}(\bar{\mathbf{a}}^t), \mathbf{S}^{*t}(\bar{\mathbf{a}}^{t-1}) : \mathbf{a}^t \in \mathcal{F}\}_{t\in\mathcal{T}}$ to be the set of potential outcomes under $\{\mathbf{a}^t\}_{t\in\mathcal{T}}$, i.e. the states and outcomes that would be observed under actions $\{\mathbf{a}^t\}_{t\in\mathcal{T}}$, where we have defined $\mathbf{S}^{*1}(\bar{\mathbf{a}}^0) \equiv \mathbf{S}^1$ for convenience.

For any $\pi \in \Pi$, let $\{\xi_\pi^t(\mathbf{s})\}_{t\in\mathcal{T},\mathbf{s}\in\mathcal{S}}$ denote a collection of independent random variables so that $P\{\xi_\pi^t(\mathbf{s}^t) = \mathbf{a}^t\} = \pi(\mathbf{a}^t; \mathbf{s}^t)$. Define

$$\mathbf{Y}^{*t}(\pi) \triangleq \sum_{\bar{\mathbf{a}}^t} \mathbf{Y}^{*t}(\bar{\mathbf{a}}^t) \prod_{v=1}^{t} \mathcal{I}[\xi_\pi^v\{\mathbf{S}^{*v}(\bar{\mathbf{a}}^{v-1})\} = \bar{\mathbf{a}}^v]$$

to be the potential outcome under allocation strategy π, where $\mathbf{S}^{*1}(\bar{\mathbf{a}}^0) = \mathbf{S}_1$. An allocation strategy $\pi^{\text{opt}} \in \Pi$ is optimal if

$$\mathbb{E}\Big[\sum_{t\geqslant 1}\gamma^{t-1}u\{\mathbf{Y}^{*t}(\pi^{\text{opt}})\}\Big] \geqslant \mathbb{E}\Big[\sum_{t\geqslant 1}\gamma^{t-1}u\{\mathbf{Y}^{*t}(\pi)\}\Big]$$

for all $\pi \in \Pi$. If there are multiple optimal strategies within Π there is no loss by choosing between them arbitrarily. Thus, for brevity, we assume hereafter that π^{opt} is unique. To estimate π^{opt} from the observed data, we require assumptions about the data-generating mechanism. At time t, the available data to estimate π^{opt} are $\mathbf{H}^1 = \mathbf{S}^1$ if $t = 1$ and $\mathbf{H}^t = (\mathbf{S}^1, \mathbf{A}^1, \mathbf{Y}^1, \ldots, \mathbf{S}^{t-1}, \mathbf{A}^{t-1}, \mathbf{Y}^{t-1}, \mathbf{S}^t)$ if $t \geqslant 2$. We make the following assumptions.

Assumption 1 (sequential ignorability (Robins, 2004)). $\mathbf{A}^t \perp\!\!\!\perp \mathbf{W}^* | \mathbf{H}^t$ for all $t \in \mathcal{T}$.

Assumption 2. The observed outcomes are the potential outcomes under treatment actually received, $\mathbf{Y}^t = \mathbf{Y}^{*t}(\bar{\mathbf{A}}^t)$ and $\mathbf{S}^t = \mathbf{S}^{*t}(\bar{\mathbf{A}}^{t-1})$ for all $t \in \mathcal{T}$.

Assumption 3 (positivity). There exists $\epsilon > 0$ so that $P(\mathbf{A}^t = \mathbf{a} | \mathbf{H}^t) > \epsilon$ for all $\mathbf{a} \in \mathcal{F}$ and $t \in \mathcal{T}$ with probability 1.

Although we have stated assumption 2 as an assumption, there is some debate about whether this should instead be taken as an axiom (Pearl, 2010; VanderWeele and Hernan, 2013; Keele, 2015); in addition, we implicitly assume throughout that there are no hidden forms of treatment. Given a data-generating process which satisfies assumptions 1–3, for any $\pi \in \Pi$ it follows that

$$\mathbb{E}\Big[\sum_{t\geqslant 1}\gamma^{t-1}u\{\mathbf{Y}^{*t}(\pi)\}\Big] = \lim_{T\to\infty}\int\Big\{\sum_{t=1}^{T}\gamma^{t-1}u(\mathbf{y}^t)\Big\}\prod_{v=1}^{T}\{f_v(\mathbf{y}^v|\mathbf{h}^v,\mathbf{a}^v)\pi(\mathbf{a}^v;\mathbf{s}^v) \times g_v(\mathbf{h}^v|\mathbf{y}^{v-1},\mathbf{h}^{v-1})\}\mathrm{d}\lambda(\bar{\mathbf{y}}^T,\bar{\mathbf{a}}^T,\bar{\mathbf{h}}^T), \quad (1)$$

where f_v is the conditional density for $\mathbf{Y}^v$ given $\mathbf{H}^v$ and $\mathbf{A}^v$, g_v is the conditional density for $\mathbf{H}^v$ given $\mathbf{Y}^{v-1}, \mathbf{H}^{v-1}$ with $g_1(\mathbf{h}^1|\mathbf{y}_0, \mathbf{h}^0) = g_1(\mathbf{h}^1)$ and λ is a dominating measure (in our applications, this will be a product of Lebesgue and counting measures corresponding to the continuous and discrete components of $\bar{\mathbf{Y}}^T$, $\bar{A}^T$ and $\bar{H}^T$). Thus, result (1) shows how the expected cumulative utility can be expressed by using the data-generating model.

The foregoing assumptions along with the assumption of no interference between experimental units are standard in causal inference for non-spatial sequential decision-making problems (Chakraborty and Moodie, 2013; Schulte *et al.*, 2014). However, in spatiotemporal decision problems, the proximity of the locations can induce spillover effects thereby causing interference between experimental units (locations) (Halloran and Struchiner, 1995; Diez Roux, 2004; Hong and Raudenbush, 2006; Hudgens and Halloran, 2008; VanderWeele and Tchetgen Tchetgen, 2011; Ogburn and VanderWeele, 2014). Furthermore, in many settings, there are cost constraints of the form $\Sigma_{l\in\mathcal{L}} v_l^t a_l^t \leqslant c_t$, where v_l^t is the cost of applying treatment at location l at time t and c_t is a total budget at time t. Constraints of this form are another reason why the decision to treat one location requires consideration of all others, i.e. applying treatment at one location reduces the available budget for applying treatments elsewhere. Standard methods for estimating optimal decision rules, e.g. Q- and A-learning (Murphy, 2003, 2005; Robins, 2004; Blatt *et al.*, 2004; Goldberg and Kosorok, 2012; Schulte *et al.*, 2014; Laber *et al.*, 2014) and policy search (Robins *et al.*, 2008; Orellana *et al.*, 2010; Zhang *et al.*, 2012a, b, 2013; Zhao *et al.*, 2012, 2014, 2015), are based on independent application of treatment to each unit. Thus, to apply these methods without additional assumptions, it would be necessary to treat the collection of all locations as a single experimental unit. In this case, there are $O(2^L)$ available allocations at each time and a single observation available for estimation. Furthermore, existing methods rely on estimation of part or all of the conditional distribution of $\mathbf{Y}^t$ given $(\bar{\mathbf{S}}^t, \bar{\mathbf{A}}^t)$ treating $\bar{\mathbf{A}}^t$ as a categorical variable with 2^L levels. Fitting such a model, even if sufficient replications were available to identify the distribution, would be computationally infeasible.

4. Estimating an optimal allocation strategy

In the context of an emerging epidemic, there is typically little or no data that can be used to form reliable estimators for some (or all) components of the system dynamics model. Thus, it is essential to add information from scientific theory to the estimation process. We integrate scientific theory with the estimation process by taking a Bayesian perspective on parameter uncertainty and allowing the use of informative priors on some (or all) of the parameters indexing our postulated system dynamics model.

An overview of our estimation procedure is as follows. Let $\mathcal{D}$ denote a class of deterministic, i.e. non-stochastic, allocation strategies. Under $d \in \mathcal{D}$, a decision maker who is presented with state $\mathbf{S} = \mathbf{s}$ will select allocation $d(\mathbf{s})$. At each time t, we draw a system dynamics model from the posterior distribution over dynamics models and subsequently use simulation–optimization (Law *et al.*, 1991; Banks, 1998; Gosavi, 2003) to compute a maximizer, say $\hat{d}^t$, of equation (1) over $\mathcal{D}$ where equation (1) is computed with respect to the sampled dynamics model. Given state $\mathbf{S}^t = \mathbf{s}^t$, the selected allocation at time t is $\hat{d}^t(\mathbf{s}^t)$. This implicitly defines a stochastic allocation, $\hat{\pi}^t(\mathbf{s}^t)$, as a mixture over $\{d(\mathbf{s}^t) : d \in \mathcal{D}\}$ with mixture probabilities equal to the posterior probability that d is the maximizer of equation (1); thus, the implied class of stochastic strategies, Π, is the class of all mixtures over strategies in $\mathcal{D}$. A schematic diagram for this procedure is displayed in Fig. 2(a).

This approach can be viewed as a version of Thompson sampling wherein allocations are

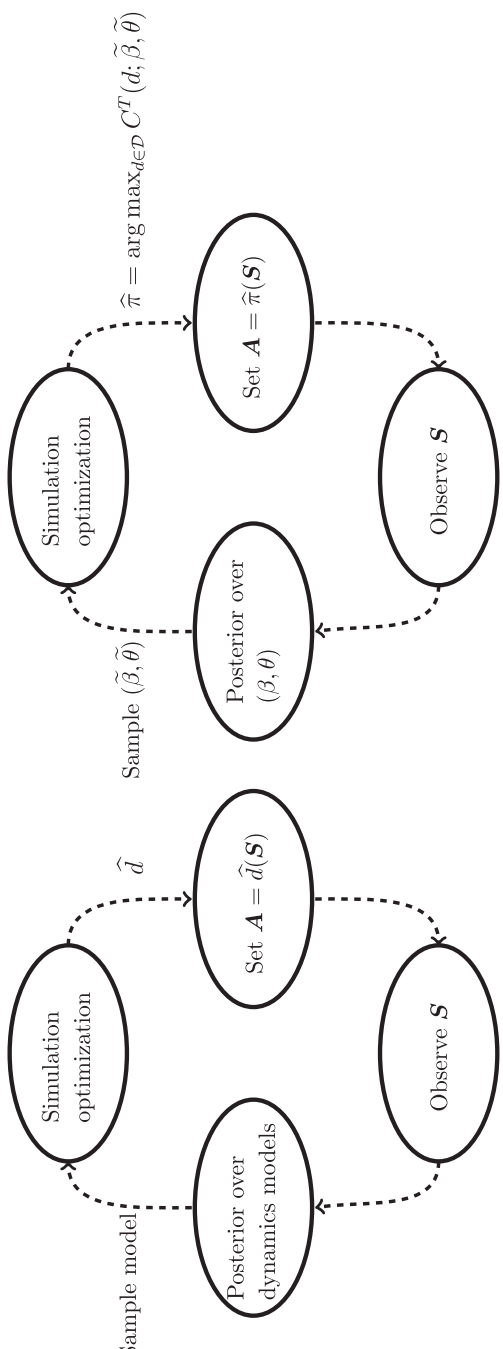

Fig. 2. (a) Schematic diagram for Thompson sampling over a generic class of dynamics models and (b) schematic diagram for Thompson sampling with a parametric class of models indexed by (θ, β) and a finite simulation horizon T

selected according to the posterior probability that they maximize the mean discounted cumulative utility (Thompson, 1933). Although Thompson sampling has been in print for more than 80 years it has only recently re-emerged in the computer science literature as a powerful tool for on-line decision making and has been shown to have several optimality properties in special cases (Chapelle and Li, 2011; Agrawal and Goyal, 2011; Kaufmann *et al.*, 2012; Korda *et al.*, 2013). Intuitively, a stochastic allocation strategy should balance exploration of the space of potential allocations with choosing allocations that are estimated to produce high expected utility; the proposed version of Thompson sampling achieves this balance through the posterior of mean utility under each $d \in \mathcal{D}$ which becomes increasingly concentrated on the maximizer as data accumulate.

To describe the implementation of our estimator we make several assumptions in addition to assumptions 1–3. We assume that the system is Markov and homogeneous in time so that, for any v, the densities in equation (1) become $f_v(\mathbf{y}^v|\mathbf{h}^v, \mathbf{a}^v) = f(\mathbf{y}^v|\mathbf{s}^v, \mathbf{a}^v)$ and $g_v(\mathbf{s}^v|\mathbf{h}^{v-1}) = g(\mathbf{s}^v|\mathbf{s}^{v-1}, \mathbf{a}^{v-1})$. Recall that the state $\mathbf{s}^v$ is a summary of the complete history up to time v, including past states, outcomes and allocations; see remark 1 for additional discussion. Furthermore, we assume parametric models $f(\mathbf{y}^v|\mathbf{s}^v, \mathbf{a}^v) = f(\mathbf{y}^v|\mathbf{s}^v, \mathbf{a}^v; \beta)$ and $g(\mathbf{s}^v|\mathbf{s}^{v-1}, \mathbf{a}^{v-1}) = g(\mathbf{s}^v|\mathbf{s}^{v-1}, \mathbf{a}^{v-1}; \theta)$ where β and θ are unknown parameters; under the model assumed, the system dynamics are completely determined by β and θ. If allocations are selected under the sequence of stochastic strategies $(\pi^1, \pi^2, \ldots, \pi^T)$, then the likelihood for (β, θ) given observed data $(\bar{\mathbf{S}}^T, \bar{\mathbf{Y}}^T, \bar{\mathbf{A}}^T)$ is

$$\mathfrak{L}_T(\beta, \theta) = \prod_{v=1}^{T} \{ f(\mathbf{Y}^v|\mathbf{S}^v, \mathbf{A}^v; \beta) \pi^v(\mathbf{A}^v; \mathbf{S}^v) g(\mathbf{S}^v|\mathbf{S}^{v-1}, \mathbf{A}^{v-1}; \theta) \},$$

where we define $g(\mathbf{s}^1|\mathbf{s}^0, \mathbf{a}^0) = g(\mathbf{s}^1)$ to be the distribution of the initial state.

For any deterministic strategy d and fixed $T > 0$, define

$$C^T(d; \beta, \theta) = \int \left\{ \sum_{t=1}^{T} \gamma^{t-1} u(\mathbf{y}^t) \right\} \prod_{v=1}^{T} [f\{\mathbf{y}^v|\mathbf{s}^v, d(\mathbf{s}^v); \beta\} g\{\mathbf{s}^v|\mathbf{s}^{v-1}, d(\mathbf{s}^{v-1}); \theta\}] \, \mathrm{d}\lambda(\bar{\mathbf{y}}^T, \bar{\mathbf{s}}^T),$$

and for $t \leqslant T$ define $\hat{\pi}^{t,T} = \arg\max_{d \in \mathcal{D}} C^T(d; \tilde{\beta}^t, \tilde{\theta}^t)$ where $\tilde{\beta}^t$ and $\tilde{\theta}^t$ are distributed according to the posterior of β, θ given $\mathbf{H}^t$. If the parametric densities are correctly specified, i.e. $f(\mathbf{y}^t|\mathbf{s}^t, \mathbf{a}^t) = f(\mathbf{y}^t|\mathbf{s}^t, \mathbf{a}^t; \beta^*)$ and $g(\mathbf{s}^t|\mathbf{s}^{t-1}, \mathbf{a}^{t-1}) = g(\mathbf{s}^t|\mathbf{s}^{t-1}, \mathbf{a}^{t-1}; \theta^*)$ for 'true' parameters β^* and θ^*, then, under standard regularity conditions (Gelman *et al.*, 2014), $\pi^{\mathrm{opt}} = \arg\max_{\pi \in \Pi} \lim_{t \to \infty} \{ \lim_{T \to \infty} C^T(d; \tilde{\beta}^t, \tilde{\theta}^t) \}$ with probability 1.

Algorithm 1 in Table 1 shows the procedure for estimating π^{opt} by using policy search with a system dynamics model and Thompson sampling; Fig. 2(b) displays a schematic diagram for this algorithm. The computational complexity of this algorithm depends on

(a) the class of strategies $\mathcal{D}$,

Table 1. Algorithm 1: policy search algorithm for an optimal allocation strategy

Input: $T < \infty, \mathbf{S}^1$
draw $\tilde{\beta}^1, \tilde{\theta}^1$ from the prior
compute $\hat{\pi}^1 = \arg\max_{d \in \mathcal{D}} C^T(d; \tilde{\beta}^1, \tilde{\theta}^1)$ (via algorithm 2 in Section 4.1)
for $j \geqslant 1$ *do*
  apply allocation $\mathbf{A}^j = \hat{\pi}^j(\mathbf{S}^j)$; observe $\mathbf{Y}^j$ and $\mathbf{S}^{j+1}$
  draw $\tilde{\beta}^{j+1}$ and $\tilde{\theta}^{j+1}$ from the posterior of (β, θ) given $\mathbf{H}^{j+1}$
  compute $\hat{\pi}^{j+1} = \arg\max_{d \in \mathcal{D}} C^T(d; \tilde{\beta}^{j+1}, \tilde{\theta}^{j+1})$ (via algorithm 2 in Section 4.1)
end

(b) the complexity of posterior distribution for β and θ, and
(c) the desired accuracy of the numerical integration that is used to compute $C^T(d;\beta,\theta)$.

In Section 4.1, we provide a class of strategies under which sampling from $\pi(\mathbf{a};\mathbf{s})$ scales linearly in the number of locations, L, making it feasible even when L is of the order of tens of thousands. In most ecological applications, the dimensions of β and θ are orders of magnitude smaller than L, e.g. the 'gravity model' for WNS (Maher *et al.*, 2012) is determined by 13 parameters; thus, integrating over the posterior of these parameters is typically not a computational bottleneck. As detailed in the next section, we use stochastic approximation to compute $\arg\max_{\pi\in\Pi} C^T(d;\beta,\theta)$; the number of Monte Carlo replicates in the numerical integration that is used to approximate $C^T(d;\beta,\theta)$ is generally smaller than L.

Remark 1. The Markov dynamics assumption that was used above is always trivially true if $\mathbf{S}^t = \mathbf{H}^t$ for all t (more formally, let $\mathbf{S}^t \in \mathbb{R}^\infty$ and define $\mathbf{S}^t = (t, \mathbf{H}^t, \mathbf{0})$, where $\mathbf{0}$ is the zero element in $\mathbb{R}^\infty$). However, this choice of state is rarely useful in large systems as the growing dimension makes modelling difficult. Thus, the Markov assumption can be viewed as an assumption about the ability of domain experts and analysts to construct a concise summary of the past that captures all salient features of the decision problem. One approach is to construct the state by concatenating information from the past k time points where k is dictated by domain knowledge or estimated from historical data. State construction for Markov decision processes is currently an active area of research (Mahadevan, 2009; Sugiyama, 2015).

Remark 2. The assumption of a low dimensional parametric model for the transition may seem overly restrictive in some settings. However, this can be relaxed through sieves (e.g. Newey (1997)) or Bayesian non-parametric models (Xu *et al.*, 2016a, b; Ghosal and van der Vaart, 2017).

4.1. A scalable class of allocation strategies

The class of allocation strategies $\mathcal{D}$ has a large influence on the quality of the estimated optimal decision strategy and the computational complexity of algorithm 1. We propose a flexible but computationally efficient class of allocation strategies that is designed to scale to large decision problems with potentially tens of thousands of locations. However, as we demonstrate in the next section, this class of strategies is also useful for problems with as few as 100 locations. Throughout, we assume that at time t exactly c^t locations can be treated; although c^t is allowed to depend on the state $\mathbf{S}^t$ we suppress this in the notation.

The proposed class of allocation strategies is based on a parametric scoring function that assigns to each location a scalar priority score with high values indicating a greater need for treatment. Because of spatial interference, the priority score at a given location must take into account the state of that location, the states of nearby locations and the configuration of treatments at nearby locations. The optimal allocation for a given scoring function assigns treatment to the locations with highest priority scores with the number of treated locations dictated by resource constraints. Each value of the parameter vector indexing the priority score corresponds to a different scoring function and hence a different optimal allocation. However, for a given value of this parameter, computing the optimal allocation requires jointly optimizing over all allowable treatment allocations, which is computationally infeasible in all except the smallest problems. Instead, for each parameter indexing the priority score, we use a greedy batch updating algorithm to approximate the optimal allocation as follows. For simplicity, assume that resource constraints allow treatment of bm locations where m is the batch size and b is the number of batches (a formal description is given below). At the first stage of the batch optimization algo-

rithm the priority scores are calculated at each location by assuming that no other locations will be treated and the m locations with highest priority scores are selected to be treated. The priority scores are then recomputed at each location by assuming that those m locations selected at the first step will be treated and the m locations with highest priority scores (required to be distinct from those selected in the first step) are selected to be treated. Then, at each subsequent step the priority scores are updated assuming that those locations selected at previous steps will be treated and the m locations with highest priority scores are selected to be treated. This procedure is applied b times until there are a total of mb locations selected for treatment. After these mb locations have been selected the procedure either terminates or can continue iterating by selecting m of the selected mb locations to be set to untreated and recomputing the priority scores and subsequently selecting m new locations for treatment. The preceding procedure requires computing a parametric score at each location for each batch update; thus, the batch size dictates how computationally expensive the updates are. In applications, the batch size can be chosen to be large initially and reduced experimentally to see whether there is any change in the solution.

The class of allocation strategies that we propose depends on a parametric class of functions from $\mathrm{supp}(\mathbf{S}^t) \times \{0,1\}^L$ into $\mathbb{R}^L$, $\mathcal{R} = \{R(\mathbf{s}^t, \mathbf{a}^t; \eta) : \eta \in E\}$, where $E \subseteq \mathbb{R}^q$. Given $\eta \in E$, the function $R(\mathbf{s}^t, \mathbf{a}^t; \eta)$ is a vector of priority scores, one per location, so $R_l(\mathbf{s}^t, \mathbf{a}^t; \eta)$ represents the priority for treating location l at time t if the observed state is $\mathbf{S}^t = \mathbf{s}^t$ and assuming that the locations $\{j : a_j^t = 1\}$ are certain to be treated. If $a_l^t = 1$ then $R_l(\mathbf{s}^t, \mathbf{a}^t; \eta) = -\infty$ so each location is selected for treatment at most once per time point. For each non-negative integer m, define the binary vector

$$\mathcal{U}_l^t(\mathbf{s}^t, \mathbf{a}^t; \eta, m) = \begin{cases} 1 & \text{if } R_l(\mathbf{s}^t, \mathbf{a}^t; \eta) \geqslant R_{(m)}(\mathbf{s}^t, \mathbf{a}^t; \eta), \\ 0 & \text{otherwise,} \end{cases}$$

where $l \in \mathcal{L}$ and $R_{(k)}(\mathbf{s}^t, \mathbf{a}^t; \eta)$ denotes the kth order statistic of $\{R_l(\mathbf{s}^t, \mathbf{a}^t; \eta)\}_{l \in \mathcal{L}}$. Let $k \leqslant c^t$ be a non-negative integer and $\mathbf{0}$ be a vector of 0s. Define $d^{(1)}(\mathbf{s}^t; \eta)$ to be the binary vector that selects the $\lfloor c^t/k \rfloor$ locations with the highest priority scores. Let $w^{(1)}$ denote $d^{(1)}(\mathbf{s}^t; \eta)$. Recursively, for $j = 2, \ldots, k$, set $w^{(j)} = d^{(j)}(\mathbf{s}^t; \eta)$, $\Delta_j = \lfloor jc^t/k \rfloor - \lfloor (j-1)c^t/k \rfloor$ and $d^{(j)}(\mathbf{s}^t, \eta) = \mathcal{U}^t\{\mathbf{s}^t, w^{(j-1)}; \eta, \Delta_j\} + w^{(j-1)}$. The final decision rule is $d(\mathbf{s}^t; \eta) = d^{(k)}(\mathbf{s}^t; \eta)$; the dependence of this rule on t occurs only through c^t.

The parameter k in the above class of strategies governs the number of locations that are selected each time that the priority scores are updated. If $k = 1$ then the priority scores are computed once, under no treatments, and the top c^t locations are treated; if $k = c^t$ then the algorithm updates the priority scores after every location selection. In large problems, we anticipate choosing $k \ll L$, e.g. $k = O\{\log(L)\}$. If the computational complexity of computing $R_l(\mathbf{s}^t, \mathbf{a}^t; \eta)$ is N, then the complexity of computing $d(\mathbf{s}^t; \eta)$ is $O(kLN)$. Thus, if $k = O\{\log(L)\}$ and N is negligible relative to L, then evaluating the strategy is $O\{L \log(L)\}$ which is feasible even for large values of L.

Let $\mathcal{D}$ denote the class of policies $\{d(\mathbf{s}; \eta) : \eta \in E\}$. Algorithm 1 requires maximization of $C^T(d; \beta, \theta)$ over $d \in \mathcal{D}$ (or, equivalently, over $\eta \in E$). Thus, the order of computation for this step depends on the complexity of evaluating the function $C^T(d; \beta, \theta)$. Obtaining a high quality approximation for $C^T(d; \beta, \theta)$ for any fixed strategy d may require a large number of expensive Monte Carlo replications; this is particularly wasteful when evaluating allocation strategies that are far from a maximizer. Thus, we use a stochastic approximation algorithm which is known as simultaneous perturbation (e.g. Spall (2005)) to approximate $\arg\max_{d \in \mathcal{D}} C^T(d; \beta, \theta)$. The algorithm relies on a sequence of non-negative step sizes $\{\alpha_j\}_{j \geqslant 1}$ and a sequence of non-negative perturbation magnitudes $\{\zeta_j\}_{j \geqslant 1}$ that satisfies $\zeta_j \to 0$ as $j \to \infty$. Convergence guarantees require that $\Sigma_{j \geqslant 1} \alpha_j = \infty$, $\Sigma_{j \geqslant 1} \alpha_j^2 < \infty$ and $\zeta_j \to 0$ as $j \to \infty$ (Kushner and Yin, 2003; Borkar, 2008);

Table 2. Algorithm 2: stochastic approximation

Input: $T<\infty$, $\mathbf{S}^t$, $\eta^0 \in E$, $f(\mathbf{y}^t|\mathbf{s}^t,\mathbf{a}^{t-1};\beta)$, $g(\mathbf{s}^t|\mathbf{s}^{t-1},\mathbf{a}^t;\theta)$, $\{\alpha_j\}_{j\geqslant 1}$, $\{\zeta_j\}_{j\geqslant 1}$ and tol >0

set $k=1$, $\tilde{\mathbf{S}}^t=\mathbf{S}^t$
do
  draw $\mathbf{Z}^k \sim \text{uniform}\{-1,1\}^d$
  for $m=0,\ldots,T-1$ *do*
    set $\mathbf{A}^{t+m}=d(\mathbf{S}^{t+m};\eta^k+\zeta_k\mathbf{Z}^k)$
    draw $\mathbf{S}^{t+m+1}\sim g(\mathbf{s}^{t+m+1}|\mathbf{S}^{t+m},\mathbf{A}^{t+m};\theta)$
    draw $\mathbf{Y}^{t+m}\sim f(\mathbf{y}^{t+m}|\mathbf{S}^{t+m},\mathbf{A}^{t+m};\beta)$
    set $\tilde{\mathbf{A}}^{t+m}=d(\tilde{\mathbf{S}}^{t+m};\eta^k-\zeta_k\mathbf{Z}^k)$
    draw $\tilde{\mathbf{Y}}^{t+m}\sim f(\mathbf{y}^{t+m}|\tilde{\mathbf{S}}^{t+m},\tilde{\mathbf{A}}^{t+m};\beta)$
    draw $\tilde{\mathbf{S}}^{t+m+1}\sim g(\mathbf{s}^{t+m+1}|\tilde{\mathbf{S}}^{t+m},\tilde{\mathbf{A}}^{t+m};\theta)$
  end
  set $\eta^{k+1}=\mathbb{G}_E[\eta^k+\{\alpha_k/(2\zeta_k)\}\{\mathbf{Z}^k\mathbf{1}_L^{\mathrm{T}}(\mathbf{Y}^{t+T-1}-\tilde{\mathbf{Y}}^{t+T-1})\}]$
  set $k=k+1$
while $\alpha_k \geqslant$ tol
Output: η^k

however, in our simulation experiments, we use $\alpha_j=\tau/(\rho+j)^{1.25}$ and $\zeta_j=100/j$, where $\tau,\rho>0$ are tuning parameters. We used a double-bootstrap procedure to select τ and ρ; details on the double-bootstrap tuning procedure are in the on-line supplemental materials. Let $\mathbb{G}_E$ denote the orthogonal projection onto E, i.e. $\mathbb{G}_E(\rho)=\arg\min_{\eta\in E}||\rho-\eta||$, where $||\cdot||$ denotes the Euclidean norm. Algorithm 2 in Table 2 shows the stochastic approximation algorithm for computing $\arg\max_{d\in\mathcal{D}} C^T(d;\beta,\theta)$; it can be seen that each iteration of this algorithm requires simulating trajectories under only two parameter values, rather than $O(d)$ parameter values as would be required by classic stochastic gradient descent methods using a difference-based approximation for the gradient (Spall, 2005; Bhatnagar *et al.*, 2013).

Remark 3. The choice of priority functions $\mathcal{R}$ will, of course, depend on features of a given application. However, for concreteness, we describe a class of linear priority functions that may be useful in practice. We use linear priority functions in our application to WNS. For each $l\in\mathcal{L}$ let $\phi_l(\mathbf{s}^t,\mathbf{a}^t)\in\mathbb{R}^p$ denote a fixed and known feature vector. Linear priority functions $\mathcal{R}_{\text{Lin}}=\{R_l(\mathbf{s}^t,\mathbf{a}^t;\eta)=\phi_l(\mathbf{s}^t,\mathbf{a}^t)^{\mathrm{T}}\eta : l\in\mathcal{L},\eta\in\mathbb{R}^p,||\eta||=1\}$ are appealing because the priority scores are interpretable and computationally simple. The features $\phi_l(\mathbf{s}^t,\mathbf{a}^t)$ can be local smooths of location covariates, e.g. measures of susceptibility or contagiousness, predictions of the disease process based on one or more postulated models for the underlying system dynamics or structural characteristics (Kolaczyk, 2009).

5. Simulation experiments

We evaluate the finite sample performance of our estimator in a suite of simulation experiments. We consider the spread of an infectious disease over two spatial domains:

(a) points in two-dimensional Euclidean space and
(b) nodes in a network.

5.1. Spread of an infectious disease in Euclidean space

In the Euclidean space setting, each location $l\in\{1,\ldots,L\}$ is a point in the unit square $[0,1]^2$.

For each location l, we generate four static covariates $\mathbf{X}_l$ by using a mean 0 Gaussian process with a multivariate separable isotropic covariance matrix that is exponential in space and autoregressive across the four covariates at each location. To mimic our motivating example of the spread of WNS, we assume that each location has a fixed number of caves; in more general spatial epidemic models, this variable represents the gravity that is associated with the location (e.g. Bossenbroek *et al.* (2001), Xia *et al.* (2004), Drake and Lodge (2004) and Sen and Smith (2012)). We generate the number of caves, Z_l, by using the first covariate and subtracting the minimum value to force non-negative values (see the on-line supplemental materials for details). The outcome Y_l^t is 1 if location l becomes infected at or before time t and 0 otherwise. The process is initialized at time $t=1$ by randomly selecting 1% of the locations to be infected. Define $\mathbf{S}_l^t = (\mathbf{X}_l, Z_l, Y_l^{t-1})$ and let $\omega_{l,k}$ denote the distance between locations l and k. We standardize the distance to have a standard deviation of 1 for computational stability. The model assumes that disease transmission is independent across locations given $(\mathbf{S}^t, \mathbf{A}^t)$, with $Y_l^t = 1$ if $Y_l^{t-1} = 1$ so that locations never shed the disease, and $P(Y_l^t = 1 | \mathbf{S}^t = \mathbf{s}^t, \mathbf{A}^t = \mathbf{a}^t) = 1 - \Pi_{k \in \mathcal{I}^t}\{1 - q_{l,k}(\mathbf{s}^t, \mathbf{a}^t)\}$ if $Y_l^{t-1} = 0$, where $\mathcal{I}^t = \{k : Y_k^{t-1} = 1\}$ is the set of sites that are infected before time t, and $q_{l,k}(\mathbf{s}^t, \mathbf{a}^t)$ is the probability that the disease spreads from location k to location l. The local dynamics are governed by the following spatial gravity model (Maher *et al.*, 2012):

$$\text{logit}\{q_{l,k}(\mathbf{s}^t, \mathbf{a}^t)\} = \theta_0 + \mathbf{x}_l^{\mathrm{T}}\boldsymbol{\theta}_1 + \mathbf{x}_k^{\mathrm{T}}\boldsymbol{\theta}_2 - \theta_3 a_l^t - \theta_4 a_k^t - \theta_5 \omega_{l,k}/(z_l z_k)^{\theta_6}, \tag{2}$$

where θ_0 is an intercept, $\boldsymbol{\theta}_1$ captures effects of the uninfected location, $\boldsymbol{\theta}_2$ captures effects of the infected location, θ_3 and θ_4 govern the strength of treatments to uninfected and infected locations, $\theta_5 > 0$ controls the spatial range of infection and $\theta_6 > 0$ controls the amount of gravity that is induced by the number of caves per location. We chose $\boldsymbol{\theta} = (\theta_0, \boldsymbol{\theta}_1^{\mathrm{T}}, \boldsymbol{\theta}_2^{\mathrm{T}}, \theta_3, \ldots, \theta_6)^{\mathrm{T}}$ to match the maximum likelihood estimator fit using data on WNS but adjusted θ_0 so that in each simulation setting an average 70% of locations become infected after $T = 15$ years in the absence of any interventions. An algorithm for constructing $\boldsymbol{\theta}$ for each simulation setting is provided in the on-line supplemental materials.

We assume that treatments can be applied to at most only ςL locations at each time t where $\varsigma \in (0, 1)$ (in all settings considered, ςL is an integer). To form a baseline for comparison, we also consider the following allocation strategies:

(a) no treatment, do not apply treatment at any location;
(b) myopic, rank uninfected locations by their estimated probability of becoming infected in the next time step, rank infected locations by the weighted average infection probability of uninfected locations by using $\lambda_{l,k}$ (defined below) as weights and then allocate treatment to the $(\varsigma/2)L$ highest ranked uninfected locations and to the $(\varsigma/2)L$ highest ranked infected locations;
(c) proximal, rank uninfected locations by their proximity (inverse distance) to the nearest infected location and rank infected locations by their proximity to uninfected locations, allocate treatment to the $(\varsigma/2)L$ highest ranked infected locations and to the $(\varsigma/2)L$ highest ranked uninfected locations;
(d) treat all locations to provide a performance ceiling if infinite resources were available and, by contrast with the no-treatment strategy, to provide a sense of the strength of treatment.

In the simulations that we present here, we set $\varsigma = 0.12$; additional simulations with other settings of ς are qualitatively similar and have thus been omitted. Because the number of possible treatment allocations is exponential in the number of locations, it is not computationally feasible to compute the optimal allocation strategy as a 'gold standard' even though the generative

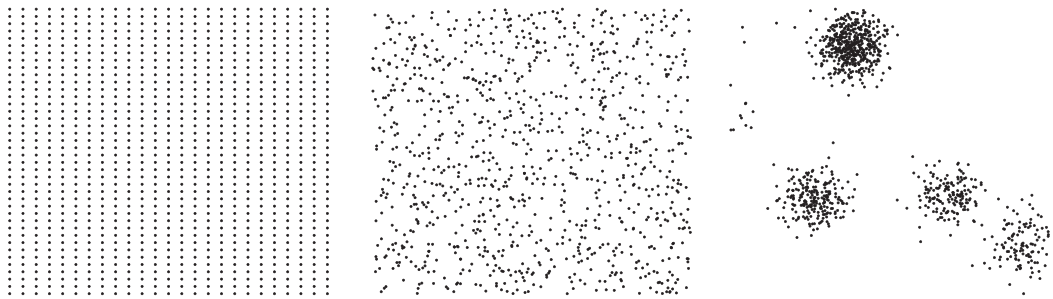

Fig. 3. (a) S1, regular lattice layout with 1000 locations, (b) S2, uniformly distributed layout with 1000 locations, and (c) S3, clustered layout with 1000 locations

distribution is known; for example in the smallest setting that we consider there are $\binom{100}{12} \approx 10^{15}$ possible allocations at each time point.

We consider three spatial location layouts: S1, a regular lattice layout, S2, a uniformly distributed layout, and S3, a clustered layout. Instances of these three layouts with $L = 1000$ locations are displayed in Fig. 3. Our simulation experiments consider location sets of size $L = 100, 500, 1000$; algorithms for generating these layouts are in the on-line supplemental materials. The infection is allowed to spread from $t = 1$ to $t = 8$ with no interventions. At time points $t = 8, \ldots, 15$, each strategy under consideration is used to choose a treatment allocation. Performance of the allocation strategies is measured in terms of the average proportion infected after $T = 15$ time points. Because there is a fixed and finite time horizon, in the implementation of our algorithm we set $\gamma = 1$.

To study the effects of model misspecification, we also consider the case when the true data-generating model is model (2) and the allocation strategies proposed are based on the postulated model (2) with $\text{logit}\{q^t_{l,k}(\mathbf{s}, \mathbf{a})\}$ replaced with $\tilde{\theta}_0 + \tilde{\theta}_1 \omega_{k,l} - \tilde{\theta}_2 a^t_l - \tilde{\theta}_3 a^t_k$ if site l is uninfected at time t and 0 otherwise. Thus, the misspecified model assumes that under no treatment the probability of spread from an infected to susceptible location is dictated by distance only.

We use linear priority functions as in remark 3: $\mathcal{R}_{\text{Lin}} = \{R_l(\mathbf{s}^t, \mathbf{a}^t; \eta) = \phi_l(\mathbf{s}^t, \mathbf{a}^t)^{\mathrm{T}} \eta : l \in \mathcal{L}, \eta \in \mathbb{R}^p, ||\eta|| = 1\}$. Let $m^t_{l,k}(\mathbf{s}^t, \mathbf{a}^t)$ denote the, possibly misspecified, postulated model for $q^t_{l,k}(\mathbf{s}^t, \mathbf{a}^t)$; write P_m to denote probabilities that are evaluated under the model postulated. Let c_l denote the half-plane data depth of location l (i.e. the minimum number of points that are contained in a hyperplane passing through l; Liu (1990)) and define $\lambda_{l,j} = \exp(\lambda \omega_{l,j}) / \Sigma_{j \neq l} \exp(\lambda \omega_{l,j})$ where the constant λ is chosen so that 80% of the total weight is placed on the $\log(L)$ nearest neighbours of location l. Recall that $\mathcal{I}^t$ denotes the set of locations that were infected before time t; let $\tilde{\mathcal{I}}^t$ denote the complement of this set. For uninfected locations, at time t, define

$$\psi_{l,1}(\mathbf{s}^t, \mathbf{a}^t) = P_m(Y^t_l = 1 | \mathbf{S}^t = \mathbf{s}^t, \mathbf{A}^t = \mathbf{a}^t),$$
$$\psi_{l,2}(\mathbf{s}^t, \mathbf{a}^t) = \psi_{l,1}(\mathbf{s}^t, \mathbf{a}^t) \sum_{j \in \mathcal{I}^t} \{1 - \psi_{j,1}(\mathbf{s}^t, \mathbf{a}^t)\} m_{l,j}(\mathbf{s}^t, \mathbf{a}^t) \lambda_{l,j},$$
$$\psi_{l,3}(\mathbf{s}^t, \mathbf{a}^t) = c_l \psi^t_{l,1}(\mathbf{s}^t, \mathbf{a}^t).$$

Thus, for uninfected locations, $\psi_{l,1}(\mathbf{s}^t, \mathbf{a}^t)$ is the probability that location l becomes infected at the next time point; $\psi_{l,2}(\mathbf{s}^t, \mathbf{a}^t)$ is the effect of treating location l on the infection probabilities of all other locations. The third feature, $\psi_{l,3}(\mathbf{s}^t, \mathbf{a}^t)$, is the probability that location l becomes infected at the next time point weighted by the data depth of location l. For infected locations, define

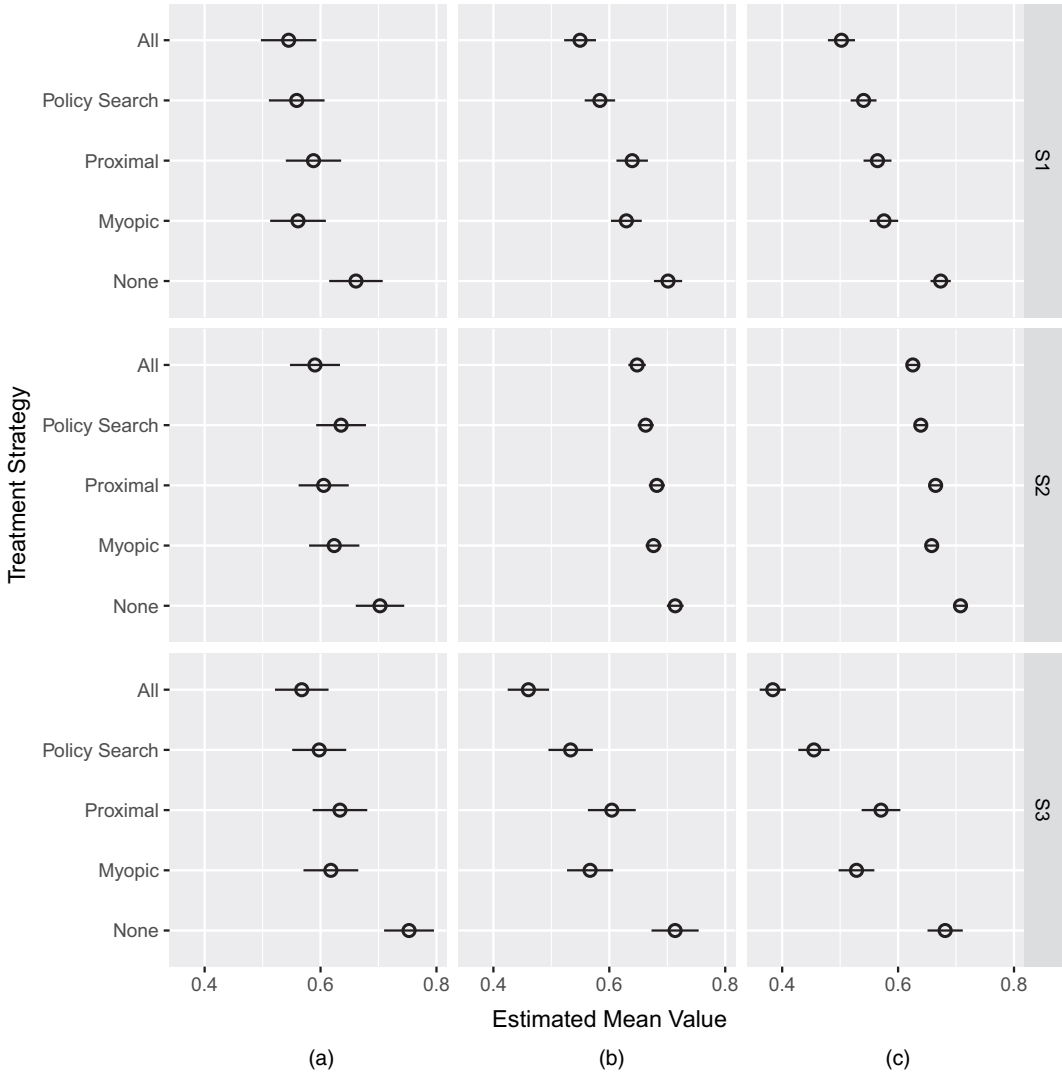

Fig. 4. Estimated average proportion infected based on 100 Monte Carlo replications under correct specification of the system dynamics model (—, 2 standard errors): (a) $L = 100$; (b) $L = 500$; (c) $L = 1000$

$$\phi_{l,1}(\mathbf{s}^t, \mathbf{a}^t) = \sum_{j \in \tilde{\mathcal{I}}^t} \lambda_{l,j} \psi_{j,1}(\mathbf{s}^t, \mathbf{a}^t),$$
$$\phi_{l,2}(\mathbf{s}^t, \mathbf{a}^t) = \sum_{j \in \tilde{\mathcal{I}}^t} \psi_{j,2}(\mathbf{s}^t, \mathbf{a}^t) m^t_{l,j}(\mathbf{s}^t, \mathbf{a}^t),$$
$$\phi_{l,3}(\mathbf{s}^t, \mathbf{a}^t) = \sum_{j \in \tilde{\mathcal{I}}^t} c_j m^t_{l,j}(\mathbf{s}^t, \mathbf{a}^t).$$

Thus, for infected locations, $\phi_{l,1}(\mathbf{s}^t, \mathbf{a}^t)$ is the weighted average infection probability over all uninfected locations at time t; $\phi_{l,2}(\mathbf{s}^t, \mathbf{a}^t)$ is a measure of expected secondary infections stemming from location l; $\phi_{l,3}(\mathbf{s}^t, \mathbf{a}^t)$ is data depth weighted by probability of infection by l.

To reduce the computational burden, we approximate the posterior distribution of $\boldsymbol{\theta}$ by using

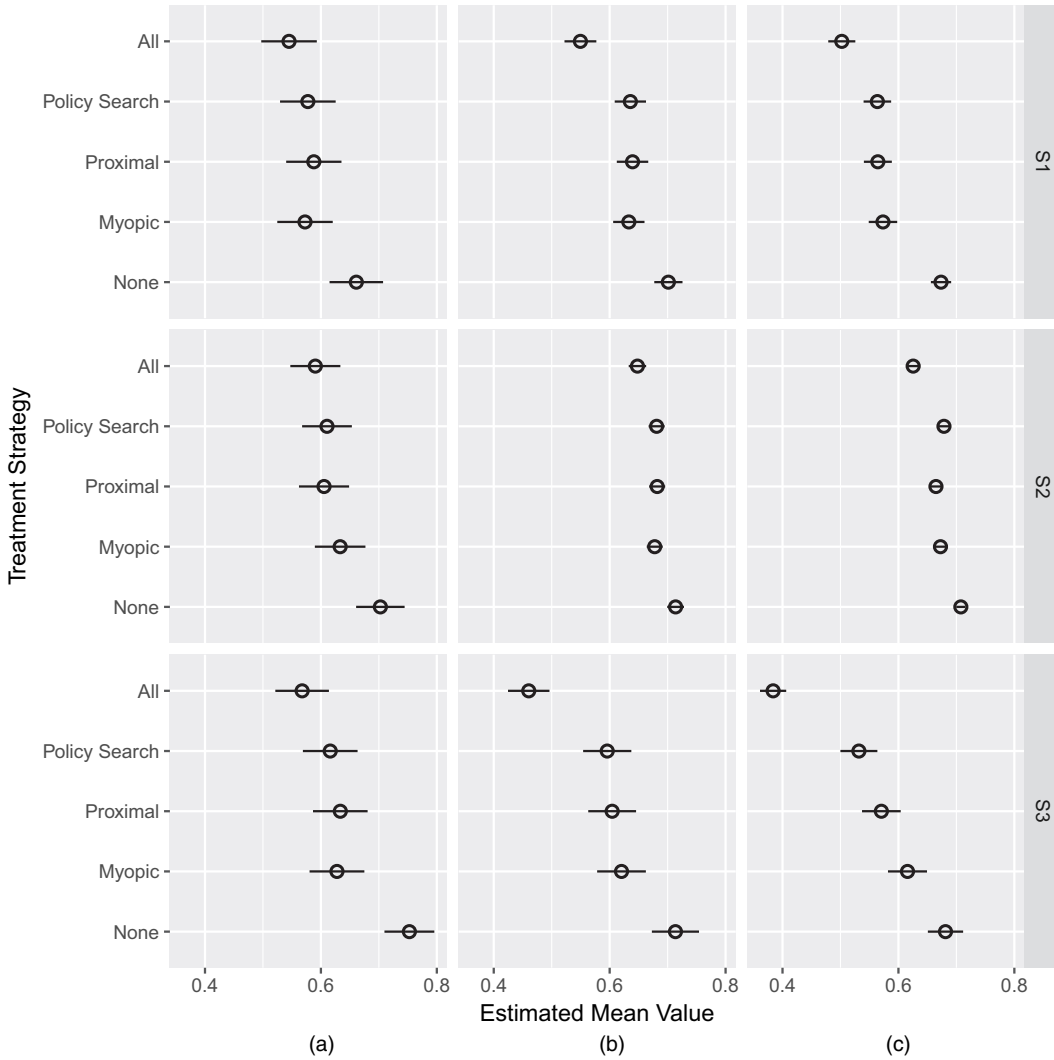

Fig. 5. Estimated average proportion infected based on 100 Monte Carlo replications under misspecification of the system dynamics model (—, 2 standard errors): (a) $L = 100$; (b) $L = 500$; (c) $L = 1000$

the plug-in estimator of the sampling distribution of the maximum likelihood estimator $\hat{\boldsymbol{\theta}}_n$; however, at the first intervention period, the treatment effects are not estimable so we sample them from the prior distribution. Additional simulations (which are not shown here) suggested that this approximation reduced the computation time by two orders of magnitude and did not affect the quality of the solution.

Simulation results under a correctly specified dynamics model are presented in Fig. 4. In all settings, the policy search algorithm proposed resulted in a smaller average proportion of infected locations than those of competing methods. Results for the misspecified dynamics model are presented in Fig. 5. As expected, the performance of policy search is worse under the incorrectly specified dynamics model; however, it still performed favourably relatively to competing methods.

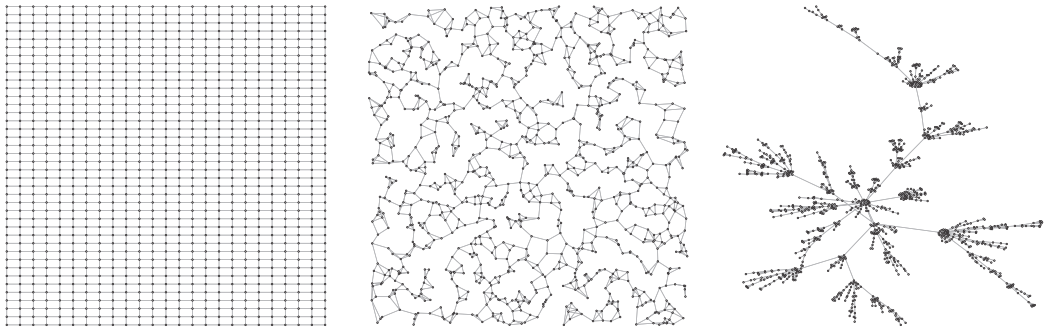

Fig. 6. (a) N1, a regular lattice network with 1000 locations, (b) N2, a random k-nearest-neighbour network with 1000 locations, and (c) N4, a small world network with 1000 locations

5.2. Spread of an infectious disease across a network

In the network setting, each location represents a node in a network with an adjacency matrix $\mathbf{\Omega}$, i.e. $\Omega_{l,k} = 1$ if locations l and k are adjacent and $\Omega_{l,k} = 0$ otherwise. Define $\mathcal{N}_l = \{k : \Omega_{l,k} = 1\}$ to be the set of adjacent locations to location l. We assume that the infection can only spread along edges in the network. Thus, if uninfected location l has zero infected neighbours at time t, so that $\mathcal{N}_l \cap \mathcal{I}_t = \varnothing$, then location l will remain uninfected at time $t+1$ with probability 1. The probability of infection is $P(Y_l^t = 1 | \mathbf{S}^t = \mathbf{s}^t, \mathbf{A}^t = \mathbf{a}^t) = 1 - \Pi_{k \in \mathcal{N}_l \cap \mathcal{I}^t}\{1 - q_{l,k}(\mathbf{s}^t, \mathbf{a}^t)\}$ where the product is now over only infected locations adjacent to location l. We define $\Pi_{k \in \varnothing} =^{\triangle} 1$ so that the probability of infection for an uninfected node is 0 when none of its neighbours are infected. The distance between any two locations is defined as the number of edges along the shortest path between them. Thus, we set $\theta_5 = 0$ in the gravity model as only neighbours of an uninfected location can influence its probability of infection. The generative model is

$$\text{logit}\{q_{l,k}(\mathbf{s}^t, \mathbf{a}^t)\} = \theta_0 + \mathbf{x}_l^{\mathrm{T}}\boldsymbol{\theta}_1 + \mathbf{x}_k^{\mathrm{T}}\boldsymbol{\theta}_2 - \theta_3 a_l^t - \theta_4 a_k^t. \tag{3}$$

As in the preceding section, we chose $\boldsymbol{\theta}$ to mimic the spread of WNS and adjusted the intercept θ_0 so that 70% of locations will be infected on average after $T = 15$ if no interventions are applied. Parameter values that were used in the simulation and details about fitting a network spread model to the observed WNS data can be found in the on-line supplementary materials. For misspecification of the system dynamics model, we remove covariate information, leaving an intercept and treatment effects. The misspecified model has the form

$$\text{logit}\{q_{l,k}(\mathbf{s}^t, \mathbf{a}^t)\} = \tilde{\theta}_0 - \tilde{\theta}_1 a_l^t - \tilde{\theta}_2 a_k^t. \tag{4}$$

We consider the following network structures: N1, a lattice, N2, a random three-nearest-neighbour graph, and N3, a small world network. Instances of these networks are displayed in Fig. 6. We use the same class of treatment strategies based on linear priority scores as in the preceding section. However, we redefine $\psi_{l,3}(\mathbf{s}^t, \mathbf{a}^t)$ and $\phi_{l,3}(\mathbf{s}^t, \mathbf{a}^t)$ to reflect distance as measured along paths in the network. In the spatial setting, we set c_l to the half-plane data depth of location l; as locations are now nodes in a network, we set c_l to be the subgraph centrality of location l (Estrada and Rodriguez-Velazquez, 2005). Additionally, because infection can spread only between adjacent nodes, we set $\lambda_{l,k} = \Omega_{l,k}$ for all l and k.

Our simulations for the spread over a network use the same competing methods and performance measures as the spread over Euclidean space. Fig. 7 shows the result when the dynamics model is correctly specified and Fig. 8 shows the results when the model is incorrectly specified.

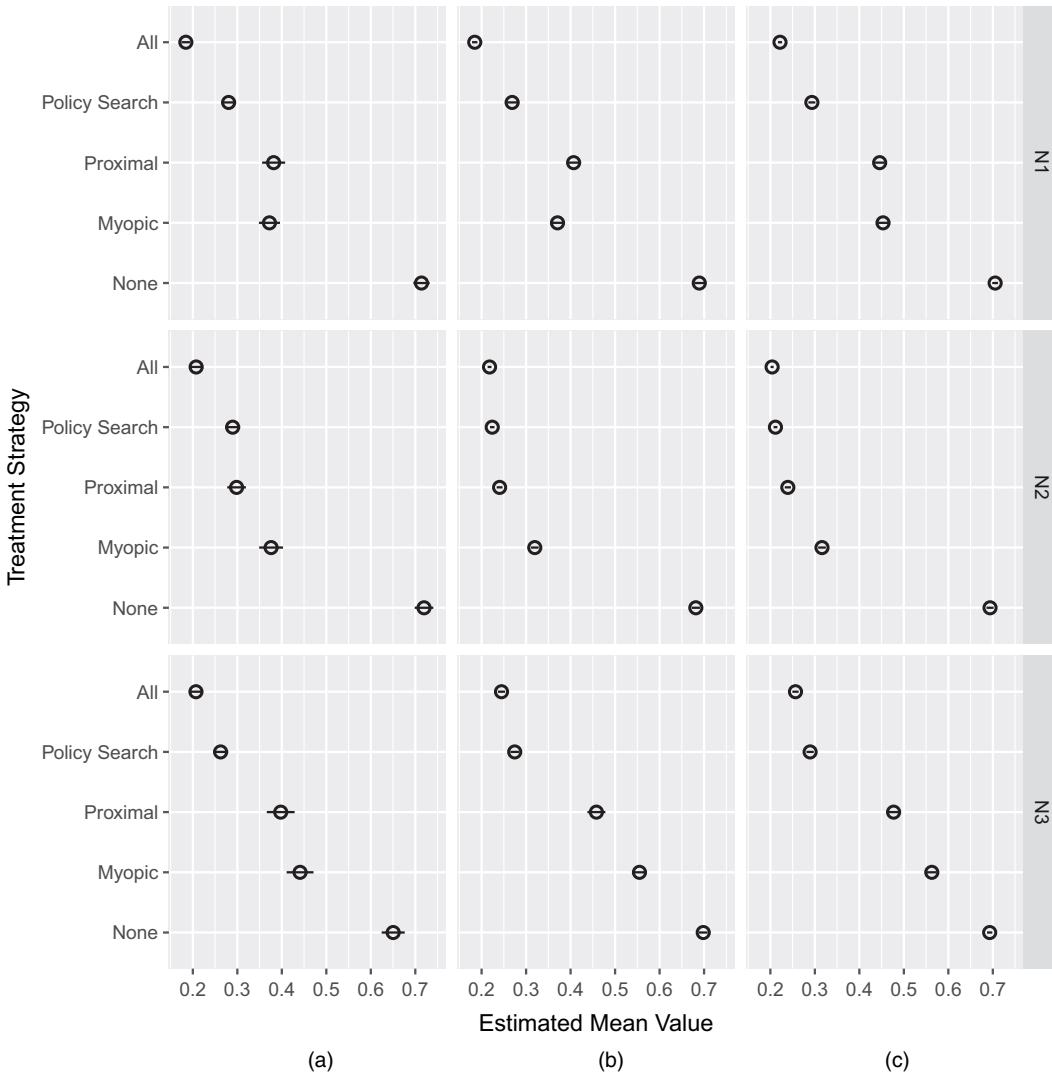

Fig. 7. Estimated average proportion infected based on 100 Monte Carlo replications under correct specification of the network spread dynamics model (—, 2 standard errors): (a) $L = 100$; (b) $L = 500$; (c) $L = 1000$

As in the preceding section, policy search performs favourably to competitors even when the dynamics model is misspecified.

6. Controlling the spread of white nose syndrome

One motivation for this work is the need to design a treatment allocation strategy to inform the management of WNS. Fig. 1 shows the current (at the time of writing) reported spread of WNS. The locations in this example are cave bearing counties that are known to house bats that are susceptible to WNS. This disease is still emerging; thus the nature of the contagion and potential treatments are still under study (e.g. Field *et al.* (2014), O'Donoghue *et al.* (2015), Turner *et al.* (2015) and Bernard *et al.* (2015)). Our goal is to use existing data on the spread

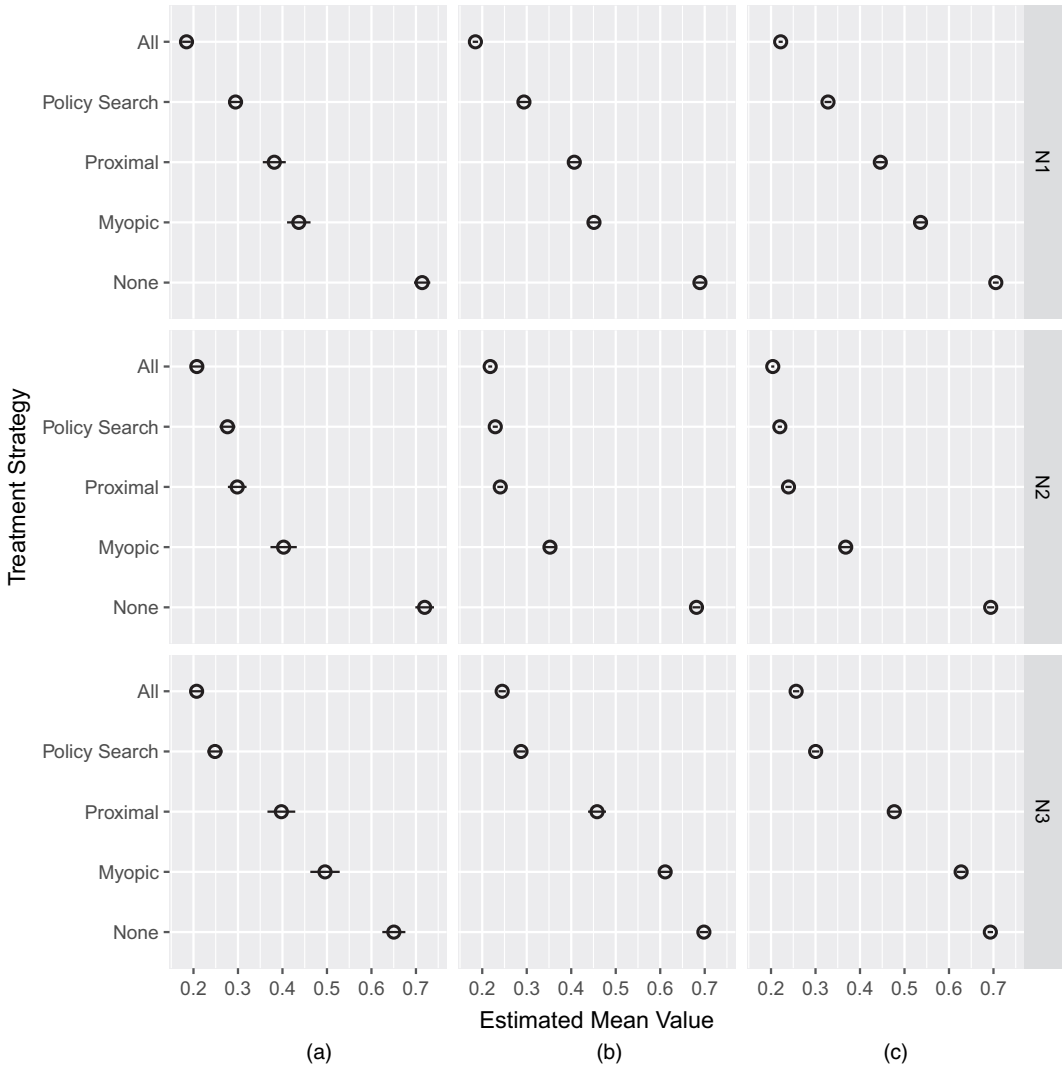

Fig. 8. Estimated average proportion infected based on 100 Monte Carlo replications under misspecification of the network spread dynamics model (—, 2 standard errors): (a) $L = 100$; (b) $L = 500$; (c) $L = 1000$

of WNS to construct an allocation strategy that could be deployed as soon as viable treatments are available. We evaluate the performance of the estimated allocation strategy by simulating the spread of WNS from 2015 to 2022 under a postulated system dynamics model.

6.1. A system dynamics model for white nose syndrome under no interventions

We begin by fitting the gravity model (2) to the WNS data that are plotted in Fig. 1. We use four static (centred and scaled) covariates in $\mathbf{X}_l$: the number of caves in the county, Z_l, the average number of days per year with temperature below 10 °C; area (in square kilometres) and species richness (the number of bat species that are thought to occupy the county). No treatments have been applied and thus $A_l^t = 0$ for all l and t, and the distance $\omega_{k,l}$ is measured

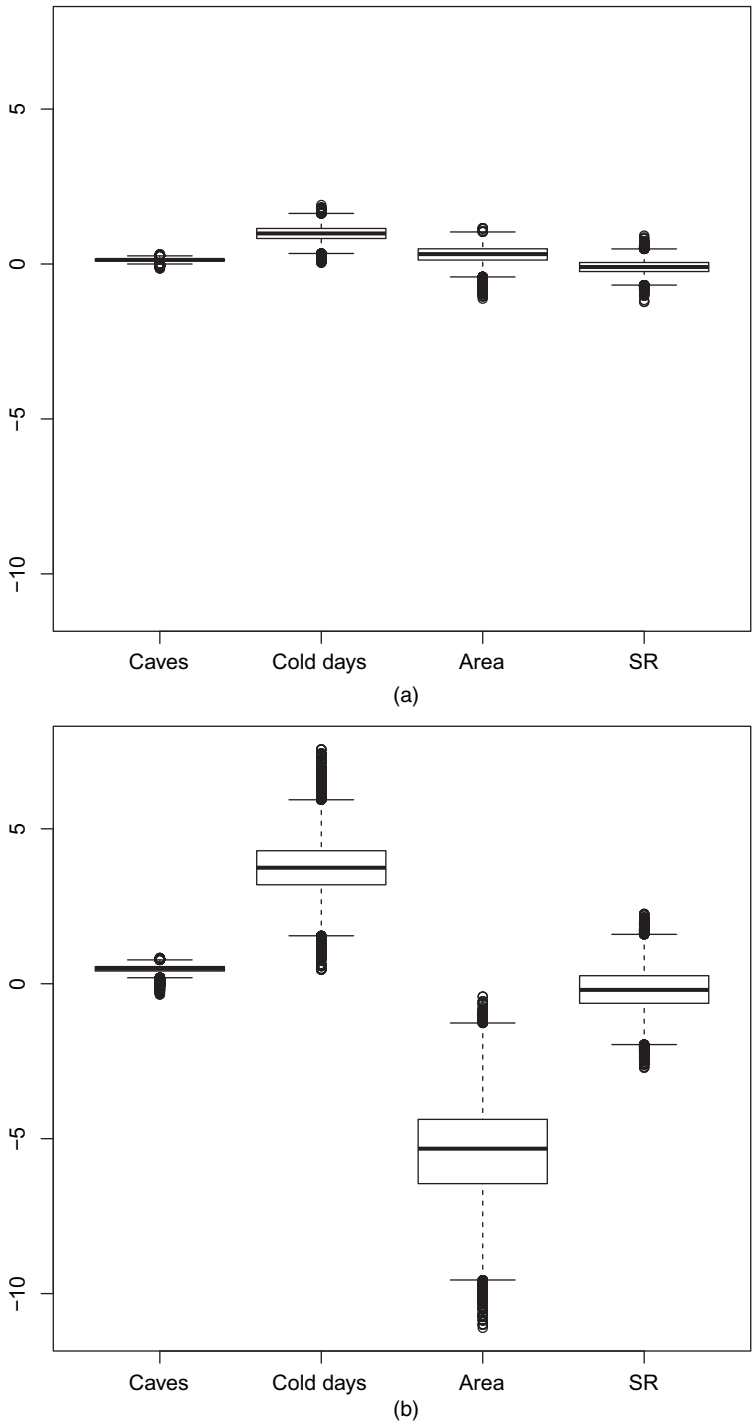

Fig. 9. Posterior distribution of the regression parameters associated with (a) uninfected (θ_1) and (b) infected (θ_2) counties in the gravity model (2) applied to the WNS data: the covariates are the number of caves in the county ('caves'), the average number of days per year below 10 °C ('cold days'), area in kilometres squared ('Area') and species richness ('SR')

Table 3. Estimated coefficients and 95% credible intervals (CIs) for the gravity model fit using WNS data from 2006 to 2014†

Parameter	*Posterior mean*	*95% credible interval*	*Value used in simulation*
θ_0	−8.35	[−10.42, −6.62]	−8.35
$\theta_{1,0}$	0.13	[0.03, 0.22]	0.13
$\theta_{1,1}$	0.99	[0.51, 1.45]	0.99
$\theta_{1,2}$	0.30	[−0.31, 0.77]	0.30
$\theta_{1,3}$	−0.10	[−0.53, 0.33]	−0.10
$\theta_{2,0}$	0.48	[0.23, 0.69]	0.48
$\theta_{2,1}$	3.76	[2.18, 5.44]	3.76
$\theta_{2,2}$	−5.45	[−8.66, −2.67]	−5.45
$\theta_{2,3}$	−0.19	[−1.53, 1.13]	−0.19
θ_3	—‡	—‡, —‡	1.77
θ_4	—‡	—‡, —‡	1.77
θ_5	24.21	[15.56, 35.86]	24.21
θ_6	0.22	[0.17, 0.28]	0.22

†Rows for θ_3 and θ_4 correspond to intervention effects which are not identifiable in the WNS data as these data do not contain any interventions.
‡Not applicable.

Table 4. Average proportion of infected counties in 100 Monte Carlo simulations of the spread of WNS from 2015 to 2022 under the gravity model†

No treatment	*Proximal*	*Myopic*	*Treat all*	*Policy search*
0.43 (0.006)	0.39 (0.006)	0.36 (0.005)	0.17 (0.001)	0.22 (0.002)

†Policy search resulted in markedly fewer infected counties than did the next best estimator.

as kilometres between county centroids. We assume an $N(0, 10^2)$ prior for θ_0 and independent $N(0, 10)$ priors for the elements of $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, and standard independent normal priors for $\log(\theta_5)$ and $\log(\theta_6)$. We sample from the posterior by using Metropolis sampling with Gaussian candidate distributions tuned to give an acceptance probability around 0.4; 100 000 samples are generated and the first 20 000 are discarded as burn-in.

The posterior is summarized in Fig. 9. As expected, the transmission probability is high when the infected and uninfected counties have many caves and many days below 10 °C. These factors increase the space and time for hibernation, and it is believed that the disease spreads primarily via contact between bats hibernating in a cave. Also, the transmission probability is high when the infected county is small, presumably because the disease can rapidly spread through the infected county and thus move quickly to nearby counties. The on-line supplemental materials include model comparisons and posterior predictive model checks which suggest that this relatively simple model is adequate for our purposes.

6.2. Simulating management of white nose syndrome

The parameter values that were used in the generative model are given in Table 3. We set $\theta_3 = \theta_4$

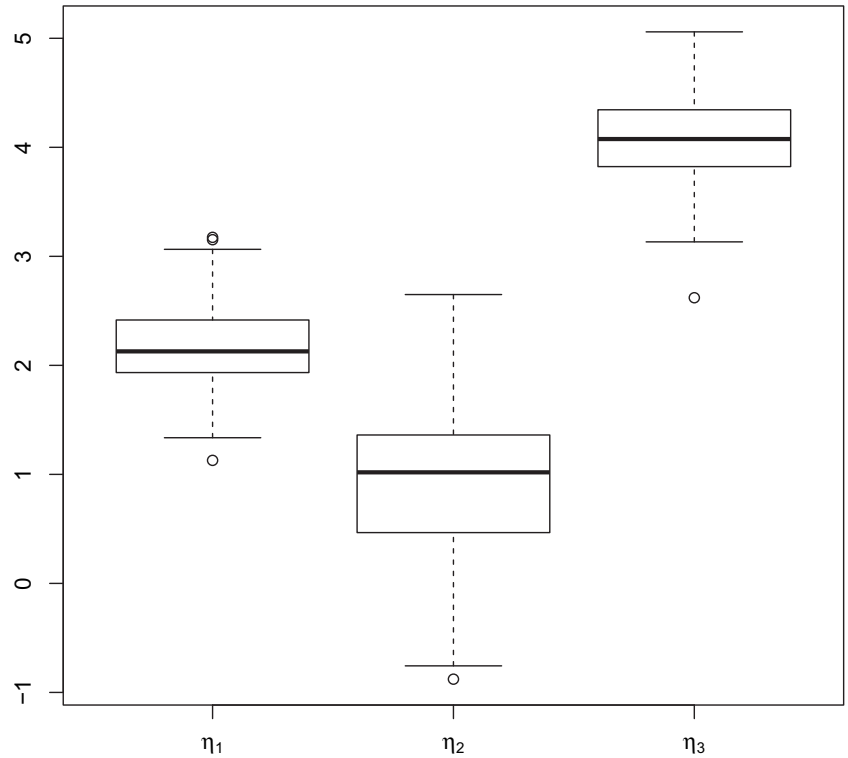

Fig. 10. Boxplots for each feature coefficient at the final time point during the WNS simulation experiment

and chose the magnitude of the treatment effect so that, if every location were treated at every time point during the management period, there would be a 95% reduction in the spread. However, when we simulate management of the disease, the strategies under consideration are limited to treating no more than 67 infected counties and 67 uninfected counties at each time point (which corresponds to treating at most 12% of the total locations). To evaluate the performance of the algorithm proposed, we simulate management of the disease from 2015 to 2022. As in Section 5, in addition to policy search, we also implement the proximal and myopic strategies. We use the same features for policy search as we did in Section 5. Also as in Section 5, in our implementation of Thompson sampling, we used draws from the estimated sampling distribution of the maximum likelihood estimator of θ to approximate draws from the posterior; at the first time point at which interventions are applied, the parameters $\theta_3 = \theta_4$ are not identified from the data so we sample them independently from an informative prior $N(4\theta_3, 1)$. Table 4 displays the average proportion of infected counties in the WNS simulation based on 100 Monte Carlo replications. The policy search algorithm resulted in significantly fewer infected counties than did competing methods. The myopic strategy is representable as in terms of the linear priority score using $\eta = (1, 0, 0)$; thus, one way to gain insight into the differences between the myopic policy and that of policy search is to examine the posterior distribution of η. Fig. 10 shows this posterior distribution. It can be seen that policy search puts large positive weights on η_3, suggesting that it is putting a high priority on the secondary effects of infection relative to the myopic strategy.

The policy search algorithm, running on an Intel Xeon Server with 64 threads (3.4 GHz; 512 Gbytes DDR4 random-access memory), took an average of 20.4 min per Monte Carlo replication. This simulation demonstrates that policy search is a feasible and potentially powerful tool that can be used to inform the management of emerging infectious diseases like WNS.

7. Discussion

We proposed a statistical framework to study sequential treatment allocation in the context of managing emerging infectious diseases. On the basis of this framework, we identified several major computational and theoretical challenges that are associated with constructing an optimal treatment allocation strategy using accumulating data on disease spread. Among these challenges is interference between locations and subsequently exponential growth in the number of allocations as a function of locations. We used a low dimensional system dynamics model to impose (implicitly) structure on the nature of treatment interference and a prespecified class of strategies to reduce the computational complexity of searching for an optimal strategy. The proposed policy search estimator performed well in simulated experiments and shows promise as a means to inform management of emerging infectious diseases.

The framework proposed used some simplifying assumptions about the nature of the spatiotemporal treatment allocation problem that could be relaxed at the expense of additional modelling and/or computational complexity. We assumed that the state $\mathbf{S}^t$ was completely observed, without error at each location at each time point. In some applications, state observations may be sparse, noisy and irregularly spaced in time. Furthermore, the sampling design may be dependent on the evolution of the disease; for example, states might be sampled only at locations where an infection had been reported. Depending on the nature of the sampling design, it may be possible to combine sampling weights or imputation methods with policy search to estimate an optimal allocations strategy. Another important simplifying assumption is that of complete compliance of the decision maker, i.e. that the recommended allocations will actually be followed. Partial compliance, which is also known as partial controllability, could be incorporated in the framework proposed by adding a compliance model that described the distribution over allocations selected by the decision maker given the allocation recommended by policy search.

As indicated in our discussion of simplifying assumptions, we believe that the area of data-driven spatiotemporal treatment allocation is rife with important and exciting open problems. We briefly review several of the most pressing of these. Our proposed estimation algorithm relies on a postulated system dynamics model; an important extension is to construct semiparametric or non-parametric estimators of the optimal allocation strategy which are robust to misspecification of the system dynamics model. One potential approach to construct such estimators is to convert the Bellman equation for the optimal allocation strategy into an estimating equation (Maei *et al.*, 2010; Ertefaie, 2014). Another important direction for additional research is scaling spatiotemporal allocation algorithms to very large problems. Our current algorithm scales readily to settings with thousands of locations; however, additional computational considerations (both in terms of central processor unit clock cycles and memory) are needed to scale to larger problems. Finally, we believe that there is the potential for rich theoretical developments regarding the difference between cumulative expected utility under an estimated allocation strategy and the optimal strategy; this difference is known as the regret in the bandit algorithm literature (Robbins, 1952). Inspired by recent theoretical developments for Thompson sampling (Korda *et al.*, 2013), we believe that it will be possible to derive minimax-type regret bounds for algorithms that are similar to our proposed policy search algorithm.

Acknowledgements

The authors gratefully acknowledge support from the National Science Foundation (grants DMS-1555141 and DMS-1513579) and the Gates Foundation.

Any use of trade, product or firms' names is for descriptive purposes only and does not imply endorsement by the US Government.

References

Agrawal, S. and Goyal, N. (2011) Analysis of Thompson sampling for the multi-armed bandit problem. *Preprint arXiv:1111.1797*.

Agrawal, S. and Goyal, N. (2012) Analysis of Thompson sampling for the multi-armed bandit problem. In *Proc. Conf. Learning Theory*, vol. 23, pp. 39.1–39.26.

Agrawal, S. and Goyal, N. (2013) Thompson sampling for contextual bandits with linear payoffs. In *Proc. Int. Conf. Machine Learning*, vol. 3, pp. 127–135.

Anderson, R. M., May, R. M. and Anderson, B. (1992) *Infectious Diseases of Humans: Dynamics and Control*. Chichester: Wiley.

Banks, J. (ed.) (1998) *Handbook of Simulation: Principles, Methodology, Advances, Applications, and Practice*. New York: Wiley.

Barbu, C., Dumonteil, E. and Gourbière, S. (2009) Optimization of control strategies for non-domiciliated triatoma dimidiata, Chagas disease vector in the Yucatán Peninsula, Mexico. *PLOS Negl. Trop. Dis.*, **3**, no. 4, article e416.

Barbu, C., Dumonteil, E. and Gourbière, S. (2011) Evaluation of spatially targeted strategies to control non-domiciliated triatoma dimidiata vector of Chagas disease. *PLOS Negl. Trop. Dis.*, **5**, no. 5, article e1045.

Bauch, C. T., Lloyd-Smith, J. O., Coffee, M. P. and Galvani, A. P. (2005) Dynamically modeling SARS and other newly emerging respiratory illnesses: past, present, and future. *Epidemiology*, **16**, 791–801.

Bernard, R. F., Foster, J. T., Willcox, E. V., Parise, K. L. and McCracken, G. F. (2015) Molecular detection of the causative agent of white-nose syndrome on Rafinesque's big-eared bats (corynorhinus rafinesquii) and two species of migratory bats in the southeastern USA. *J. Wldlif. Dis.*, **51**, 519–522.

Berry, D. A. and Fristedt, B. (1985) *Bandit Problems: Sequential Allocation of Experiments*. New York: Springer.

Bertsekas, D. P. (1996) *Neuro-dynamic Programming*. Nashua: Athena Scientific.

Bhatnagar, S., Prasad, H. and Prashanth, L. (2013) Kiefer-Wolfowitz algorithm. In *Stochastic Recursive Algorithms for Optimization*, pp. 31–39. New York: Springer.

Blatt, D., Murphy, S. A. and Zhu, J. (2004) A-learning for approximate planning. *Technical Report 04-63*, Methodology Center, Pennsylvania State University, State College.

Blehert, D. S., Hicks, A. C., Behr, M., Meteyer, C. U., Berlowski-Zier, B. M., Buckles, E. L., Coleman, J. T., Darling, S. R., Gargas, A. and Niver, R. (2009) Bat white-nose syndrome: an emerging fungal pathogen? *Science*, **323,** 227.

Borkar, V. S. (2008) *Stochastic Approximation*, vol. 1. New York: Cambridge University Press.

Bossenbroek, J. M., Kraft, C. E. and Nekola, J. C. (2001) Prediction of long-distance dispersal using gravity models: zebra mussel invasion of inland lakes. *Ecol. Appl.*, **11**, 1778–1788.

Boyles, J. G., Cryan, P. M., McCracken, G. F. and Kunz, T. H. (2011) Economic importance of bats in agriculture. *Science*, **332**, 41–42.

Bozzette, S. A., Boer, R., Bhatnagar, V., Brower, J. L., Keeler, E. B., Morton, S. C. and Stoto, M. A. (2003) A model for a smallpox-vaccination policy. *New Engl. J. Med.*, **348**, 416–425.

Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A. and Wiener, J. (2000) Graph structure in the web. *Comput. Netwrks*, **33**, 309–320.

Cesa-Bianchi, N. and Lugosi, G. (2006) *Prediction, Learning, and Games*. Cambridge: Cambridge University Press.

Chakraborty, B. and Moodie, E. E. (2013) *Statistical Methods for Dynamic Treatment Regimes*. New York: Springer.

Chapelle, O. and Li, L. (2011) An empirical evaluation of Thompson sampling. In *Advances in Neural Information Processing Systems*, pp. 2249–2257.

Cornelison, C. T., Keel, M. K., Gabriel, K. T., Barlament, C. K., Tucker, T. A., Pierce, G. E. and Crow, S. A. (2014) A preliminary report on the contact-independent antagonism of pseudogymnoascus destructans by rhodococcus rhodochrous strain dap96253. *BMC Micrbiol.*, **14**, no. 1, article 246.

Diez Roux, A. V. (2004) Estimating neighborhood health effects: the challenges of causal inference in a complex world. *Socl Sci. Med.*, **58**, 1953–1960.

Drake, J. M. and Lodge, D. M. (2004) Global hot spots of biological invasions: evaluating options for ballast–water management. *Proc. R. Soc. Lond.* B, **271**, 575–580.

Ertefaie, A. (2014) Constructing dynamic treatment regimes in infinite-horizon settings. *Preprint arXiv:1406.0764*.

Estrada, E. and Rodriguez-Velazquez, J. A. (2005) Subgraph centrality in complex networks. *Phys. Rev.* E, **71**, no. 5, article 056103.
Ferguson, N. M., Donnelly, C. A. and Anderson, R. M. (2001a) The foot-and-mouth epidemic in Great Britain: pattern of spread and impact of interventions. *Science*, **292**, 1155–1160.
Ferguson, N. M., Donnelly, C. A., and Anderson, R. M. (2001b) Transmission intensity and impact of control policies on the foot and mouth epidemic in Great Britain. *Nature*, **413**, 542–548.
Ferguson, N. M., Keeling, M. J., Edmunds, W. J., Gani, R., Grenfell, B. T., Anderson, R. M. and Leach, S. (2003) Planning for smallpox outbreaks. *Nature*, **425**, 681–685.
Field, K., Reeder, S., Rogers, E., James, M., Sigler, L., Vodzak, M. Moore, M., Johnson, J. and Reeder, D. (2014) Anti-fungal immune responses to pseudogymnoasces destructans in bats affected by white-nose syndrome (vet2p. 1041). *J. Immun.*, **192**, suppl. 1, 207–213.
Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2014) *Bayesian Data Analysis*, vol. 2. New York: Taylor and Francis.
Ghavamzadeh, M., Mannor, S., Pineau, J. and Tamar, A. (2015) *Bayesian Reinforcement Learning: a Survey*. Singapore: World Scientific.
Ghosal, S. and van der Vaart, A. (2017) *Fundamentals of Nonparametric Bayesian Inference*. New York: Cambridge University Press.
Goldberg, Y. and Kosorok, M. R. (2012) Q-learning with censored data. *Ann. Statist.*, **40**, 529–560.
Gopalan, A. and Mannor, S. (2015) Thompson sampling for learning parameterized Markov decision processes. In *Proc. 28th Conf. Learning Theory*, pp. 861–898.
Gopalan, A., Mannor, S. and Mansour, Y. (2014) Thompson sampling for complex online problems. In *Proc. Int. Conf. Machine Learning*, vol. 14, pp. 100–108.
Gosavi, A. (2003) *Simulation-based Optimization: Parametric Optimization Techniques and Reinforcement Learning*. New York: Springer.
Halloran, M. E. and Struchiner, C. J. (1995) Causal inference in infectious diseases. *Epidemiology*, **6**, 142–151.
Hollingsworth, T. D. (2009) Controlling infectious disease outbreaks: lessons from mathematical modelling. *J. Publ. Hlth Poly*, **30**, 328–341.
Hollingsworth, T. D., Ferguson, N. M. and Anderson, R. M. (2006) Will travel restrictions control the international spread of pandemic influenza? *Nat. Med.*, **12**, 497–499.
Hong, G. and Raudenbush, S. W. (2006) Evaluating kindergarten retention policy. *J. Am. Statist. Ass.*, **101**, 901–910.
Hoyt, J. R., Cheng, T. L., Langwig, K. E., Hee, M. M., Frick, W. F. and Kilpatrick, A. M. (2015) Bacteria isolated from bats inhibit the growth of pseudogymnoascus destructans, the causative agent of white-nose syndrome. *PLOS One*, **10**, no. 4, article e0121329.
Huang, C.-Y., Sun, C.-T., Hsieh, J.-L. and Lin, H. (2004) Simulating SARS: small-world epidemiological modeling and public health policy assessments. *J. Artif. Soc. Socl Simuln*, **7**, no. 4, article 2.
Hudgens, M. G. and Halloran, M. E. (2008) Toward causal inference with interference. *J. Am. Statist. Ass.*, **103**, 832–842.
Hufnagel, L., Brockmann, D. and Geisel, T. (2004) Forecast and control of epidemics in a globalized world. *Proc. Natn. Acad. Sci. USA*, **101**, 15124–15129.
Jacquez, J. A., Simon, C. P., Koopman, J., Sattenspiel, L. and Perry, T. (1988) Modeling and analyzing HIV transmission: the effect of contact patterns. *Math. Biosci.*, **92**, 119–199.
Jung, E., Iwami, S., Takeuchi, Y. and Jo, T.-C. (2009) Optimal control strategy for prevention of avian influenza pandemic. *J. Theoret. Biol.*, **260**, 220–229.
Kaelbling, L. P., Littman, M. L. and Moore, A. W. (1996) Reinforcement learning: a survey. *J. Artif. Intell. Res.*, **4**, 237–285.
Kang, C., Janes, H. and Huang, Y. (2014) Combining biomarkers to optimize patient treatment recommendations. *Biometrics*, **70**, 695–707.
Kaplan, E. H., Craft, D. L. and Wein, L. M. (2002) Emergency response to a smallpox attack: the case for mass vaccination. *Proc. Natn. Acad. Sci. USA*, **99**, 10935–10940.
Kaufmann, E., Korda, N. and Munos, R. (2012) Thompson sampling: an asymptotically optimal finite-time analysis. In *Algorithmic Learning Theory* (eds N. H. Bshouty, G. Stoltz, N. Vayatis and T. Zeugmann), pp. 199–213. New York: Springer.
Keele, L. (2015) The statistics of causal inference: a view from political methodology. *Polit. Anal.*, **23**, 313–335.
Keeling, M. J. (2005) Models of foot-and-mouth disease. *Proc. R. Soc. Lond.* B, **272**, 1195–1202.
Keeling, M. J. and Rohani, P. (2011) *Modeling Infectious Diseases in Humans and Animals*. Princeton: Princeton University Press.
Kilpatrick, A. M. (2011) Globalization, land use, and the invasion of West Nile virus. *Science*, **334**, 323–327.
Kohn, K. W. (1999) Molecular interaction map of the mammalian cell cycle control and DNA repair systems. *Molec. Biol. Cell*, **10**, 2703–2734.
Kolaczyk, E. D. (2009) *Statistical Analysis of Network Data*. New York: Springer.
Korda, N., Kaufmann, E. and Munos, R. (2013) Thompson sampling for 1-dimensional exponential family bandits. In *Advances in Neural Information Processing Systems*, pp. 1448–1456.

Korenromp, E. L., Van Vliet, C., Grosskurth, H., Gavyole, A., Van der Ploeg, C. P., Fransen, L., Hayes, R. J. and Habbema, J. D. F. (2000) Model-based evaluation of single-round mass treatment of sexually transmitted diseases for HIV control in a rural African population. *Aids*, **14**, 573–593.
Kramer, A. M., Pulliam, J. T., Alexander, L. W., Park, A. W., Rohani, P. and Drake, J. M. (2016) Spatial spread of the West Africa Ebola epidemic. *Open Sci.*, **3**, no. 8, article 160294.
Kretzschmar, M., Van den Hof, S., Wallinga, J. and Van Wijngaarden, J. (2004) Ring vaccination and smallpox control. *Emergng Infect. Dis.*, **10**, no. 5, 832–841.
Kushner, H. J. and Yin, G. (2003) *Stochastic Approximation and Recursive Algorithms and Applications*. New York: Springer.
Laber, E. B., Linn, K. A. and Stefanski, L. A. (2014) Interactive model building for q-learning. *Biometrika*, **101**, 831–847.
Laber, E. and Zhao, Y. (2015) Tree-based methods for estimating individualized treatment regimes. *Biometrika*, **102**, 501–514.
Law, A. M. and Kelton, W. D. (1991) *Simulation Modeling and Analysis*, vol. 2. New York: McGraw-Hill.
Lekone, P. E. and Finkenstädt, B. F. (2006) Statistical inference in a stochastic epidemic SEIR model with control intervention: Ebola as a case study. *Biometrics*, **62**, 1170–1177.
Le Menach, A., Vergu, E., Grais, R. F., Smith, D. L. and Flahault, A. (2006) Key strategies for reducing spread of avian influenza among commercial poultry holdings: lessons for transmission to humans. *Proc. R. Soc. Lond.* B, **273**, 2467–2475.
Li, S.-L., Bjørnstad, O. N., Ferrari, M. J., Mummah, R., Runge, M. C., Fonnesbeck, C. J., Tildesley, M. J., Probert, W. J. and Shea, K. (2017) Essential information: uncertainty and optimal control of Ebola outbreaks. *Proc. Natn. Acad. Sci. USA*, **114**, 5659–5664.
Liu, R. Y. (1990) On a notion of data depth based on random simplices. *Ann. Statist.*, **18**, 405–414.
Ma, Z., Zhou, Y. and Wu, J. (2009) *Modeling and Dynamics of Infectious Diseases*, vol. 11. Singapore: World Scientific Publishers.
Maei, H. R., Szepesvári, C., Bhatnagar, S. and Sutton, R. S. (2010) Toward off-policy learning control with function approximation. In *Proc. 27th Int. Conf. Machine Learning*, pp. 719–726.
Mahadevan, S. (2009) *Learning Representation and Control in Markov Decision Processes*. Breda: Now.
Maher, S. P., Kramer, A. M., Pulliam, J. T., Zokan, M. A., Bowden, S. E., Barton, H. D., Magori, K. and Drake, J. M. (2012) Spread of white-nose syndrome on a network regulated by geography and climate. *Nat. Communs*, **3**, article 1306.
Murphy, S. A. (2003) Optimal dynamic treatment regimes (with discussion). *J. R. Statist. Soc.* B, **65**, 331–366.
Murphy, S. A. (2005) A generalization error for Q-learning. *J. Mach. Learn. Res.*, **6**, 1073–1097.
Newey, W. K. (1997) Convergence rates and asymptotic normality for series estimators. *J. Econmetr.*, **79**, 147–168.
Neyman, J. (1990) On the application of probability theory to agricultural experiments: essay on principles, section 9 (Engl. transl. D. M. Dabrowska and T. P. Speed). *Statist. Sci.*, **5**, 465–472.
O'Donoghue, A. J., Knudsen, G. M., Beekman, C., Perry, J. A., Johnson, A. D., DeRisi, J. L., Craik, C. S. and Bennett, R. J. (2015) Destructin-1 is a collagen-degrading endopeptidase secreted by pseudogymnoascus destructans, the causative agent of white-nose syndrome. *Proc. Natn. Acad. Sci. USA*, **112**, 7478–7483.
Ogburn, E. L. and VanderWeele, T. J. (2014) Vaccines, contagion, and social networks. *Preprint arXiv:1403.1241*.
Orellana, L., Rotnitzky, A. and Robins, J. (2010) Dynamic regime marginal structural mean models for estimation of optimal dynamic treatment regimes, part i: main content. *Int. J. Biostatist.*, **6**, no. 2, 1–49.
Osband, I., Russo, D. and Van Roy, B. (2013) (More) efficient reinforcement learning via posterior sampling. In *Advances in Neural Information Processing Systems*, pp. 3003–3011.
Pearl, J. (2010) On the consistency rule in causal inference: axiom, definition, assumption, or theorem? *Epidemiology*, **21**, 872–875.
Powell, W. B. (2007) *Approximate Dynamic Programming: Solving the Curses of Dimensionality*. New York: Wiley.
Rich, B. (2013) Optimal dynamic treatment regime structural nested mean models: improving efficiency through diagnostics and re-weighting and application to adaptive individualized dosing. *Dissertation*. McGill University, Montreal.
Rich, B., Moodie, E. E. M., Stephens, D. A. and Platt, R. W. (2014) Simulating sequential multiple assignment randomized trials to generate optimal personalized warfarin dosing strategies. *Clin. Trials*, **11**, 435–444.
Riley, S. (2007) Large-scale spatial-transmission models of infectious disease. *Science*, **316**, 1298–1301.
Robbins, H. (1952) Some aspects of the sequential design of experiments. *Bull. Am. Math. Soc.*, **58**, 527–535.
Robins, J. M. (2004) Optimal structural nested models for optimal sequential decisions. In *Proc. 2nd Seattle Symp. Biostatistics*, pp. 189–326. New York: Springer.
Robins, J., Orellana, L. and Rotnitzky, A. (2008) Estimation and extrapolation of optimal treatment and testing strategies. *Statist. Med.*, **27**, 4678–4721.
Rubin, D. (1978) Bayesian inference for causal effects: the role of randomization. *Ann. Statist.*, **6**, 34–58.
Russo, D. and Van Roy, B. (2014) An information-theoretic analysis of Thompson sampling. *J. Mach. Learn. Res.*, **17**, 1–30.
Schulte, P., Tsiatis, A., Laber, E. and Davidian, M. (2014) Q- and a-learning methods for estimating optimal dynamic treatment regimes. *Statist. Sci.*, **29**, 640–661.

Scott, S. L. (2010) A modern bayesian look at the multi-armed bandit. *Appl. Stochast. Modls Bus. Indstry*, **26**, 639–658.
Sen, A. and Smith, T. (2012) *Gravity Models of Spatial Interaction Behavior*. New York: Springer Science and Business Media.
Sobel, M. E. (2006) What do randomized studies of housing mobility demonstrate?: Causal inference in the face of interference. *J. Am. Statist. Ass.*, **101**, 1398–1407.
Spall, J. C. (2005) *Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control*. New York: Wiley.
Strogatz, S. H. (2001) Exploring complex networks. *Nature*, **410**, 268–276.
Subcommittee on Fisheries, Wildlife, and Oceans (2011) *Why We Should Care about Bats: Devastating Impact White-nose Syndrome is having on One of Nature's Best Pest Controllers*. Washington DC: US Government Printing Office.
Sugiyama, M. (2015) *Statistical Reinforcement Learning: Modern Machine Learning Approaches*. Boca Raton: CRC Press.
Sutton, R. and Barto, A. (1998) *Reinforcment Learning: an Introduction*. Cambridge: Massachusetts Institute of Technology Press.
Szymanski, J. A., Runge, M. C., Parkin, M. J. and Armstrong, M. (2009) White-nose syndrome management: report on structured decision making initiative. *Report*. Department of the Interior, US Fish and Wildlife Service, Fort Snelling.
Thompson, W. R. (1933) On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, **25**, 285–294.
Tildesley, M. J., Savill, N. J., Shaw, D. J., Deardon, R., Brooks, S. P., Woolhouse, M. E., Grenfell, B. T. and Keeling, M. J. (2006) Optimal reactive vaccination strategies for a foot-and-mouth outbreak in the UK. *Nature*, **440**, 83–86.
Truscott, J. and Ferguson, N. M. (2012) Evaluating the adequacy of gravity models as a description of human mobility for epidemic modelling. *PLOS Computnl Biol.*, **8**, no. 10, article e1002699.
Turner, J. M., Warnecke, L., Wilcox, A., Baloun, D., Bollinger, T. K., Misra, V. and Willis, C. K. (2015) Conspecific disturbance contributes to altered hibernation patterns in bats with white-nose syndrome. *Physiol. Behav.*, **140**, 71–78.
US Fish and Wildlife Service (2015) White-nose syndrome: a coordinated response to the devastating bat disease. US Fish and Wildlife Service, Fort Snelling.
VanderWeele, T. J. and Hernan, M. A. (2013) Causal inference under multiple versions of treatment. *J Causl Inf.*, **1**, 1–20.
VanderWeele, T. J. and Tchetgen Tchetgen, E. J. (2011) Effect partitioning under interference in two-stage randomized vaccine trials. *Statist. Probab. Lett.*, **81**, 861–869.
Williams, R. J. and Martinez, N. D. (2000) Simple rules yield complex food webs. *Nature*, **404**, 180–183.
Xia, Y., Bjørnstad, O. N. and Grenfell, B. T. (2004) Measles metapopulation dynamics: a gravity model for epidemiological coupling and dynamics. *Am. Natlst*, **164**, 267–281.
Xu, X., Kypraios, T. and O'Neill, P. D. (2016a) Bayesian non-parametric inference for stochastic epidemic models using Gaussian processes. *Biostatistics*, **17**, 619–633.
Xu, Y., Müller, P., Wahed, A. S. and Thall, P. F. (2016b) Bayesian nonparametric estimation for dynamic treatment regimes with sequential transition times. *J. Am. Statist. Ass.*, **111**, 921–950.
Zhang, B., Tsiatis, A. A., Davidian, M., Zhang, M. and Laber, E. (2012a) Estimating optimal treatment regimes from a classification perspective. *Stat*, **1**, 103–114.
Zhang, B., Tsiatis, A. A., Laber, E. B. and Davidian, M. (2012b) A robust method for estimating optimal treatment regimes. *Biometrics*, **68**, 1010–1018.
Zhang, B., Tsiatis, A. A., Laber, E. B. and Davidian, M. (2013) Robust estimation of optimal dynamic treatment regimes for sequential treatment decisions. *Biometrika*, **100**, 681–694.
Zhao, Y.-Q., Zeng, D., Laber, E. B. and Kosorok, M. R. (2015) New statistical learning methods for estimating optimal dynamic treatment regimes. *J. Am. Statist. Ass.*, **110**, 583–598.
Zhao, Y.-Q., Zeng, D., Laber, E. B., Song, R., Yuan, M. and Kosorok, M. R. (2014) Doubly robust learning for estimating individualized treatment with censored data. *Biometrika*, **102**, 151–168.
Zhao, Y., Zeng, D., Rush, A. J. and Kosorok, M. R. (2012) Estimating individualized treatment rules using outcome weighted learning. *J. Am. Statist. Ass.*, **107**, 1106–1118.

Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

'Supplementary material'.

Discussion on the paper by Laber, Meyer, Reich, Pacifici, Collazo and Drake

Theodore Kypraios (*University of Nottingham*)

I congratulate the authors for a thought-provoking paper and I am delighted to open the discussion.

Understanding the spread of communicable infectious diseases is of great importance to prevent major future outbreaks. Therefore, enormous attention has been devoted to the development of methods for fitting models to data; see, for example, O'Neill (2010) for a review. However, to control the spread of an outbreak *while it is in progress* we need to know where and when to apply any interventions.

It has already been demonstrated in the literature that, once a model has been fitted to the available data to date, then realizations of the (fitted) model can be simulated and hence provide us with information about how the epidemic might unfold in the future (see Jewell *et al.* (2009a), for example). However, in my opinion, the key contribution of this paper is that it introduces a *formal* framework to inform such decisions (in realtime) by formalizing a treatment allocation strategy. Although the methodology proposed is closely related to the idea of dynamic treatment regimes in personalized medicine, there are several challenges, such as scarcity of data, scalability and dependence between locations, that prevent the latter's direct application to the problem of spatiotemporal treatment allocation. However, the authors have successfully addressed these challenges and opened up new avenues for further research.

I shall focus my discussion around the modelling assumptions of the system dynamics. The authors consider a spatial gravity model (Maher *et al.*, 2012) which has a generalized linear model flavour and is concerned with modelling the probability that the disease spreads from location k to location l at time t. Given observed data up to time t the model can be fitted by using either maximum likelihood or within a Bayesian framework. However, it is natural to model the spread of infectious diseases by using mechanistic (deterministic or stochastic) models whose key aspect is the rate at which new infections occur and how long a unit remain infectious for. Therefore a natural question of interest is how easy and/or practical is it to adapt the proposed methodology if we were to consider a (heterogeneously mixing) stochastic susceptible–infected–removed (SIR) model, such as that, for example, in Jewell *et al.* (2009b)?

In the white nose syndrome in bats application, the authors assume that, for each county, the infection status and time since infected as well as other covariates (the number of caves, average winter temperature etc.) are observed. But let us consider instead modelling a foot-and-mouth outbreak in UK farms by using a heterogeneously mixing SIR model (Jewell *et al.*, 2009b; Deardon *et al.*, 2010). The framework developed appears ideal to guide the design of control strategies during the course of an outbreak. However, in contrast with the white nose syndrome application, at any particular time during the outbreak,

(a) there might be farms that are infected but we have not noticed yet and
(b) the times at which infected farms were infected are not observed.

In consequence the likelihood of the observed data (e.g. culling dates and the onset symptoms) is intractable and, often, computationally intensive methods are employed to make inference for the model parameters (Jewell *et al.*, 2009b; Deardon *et al.*, 2010). Would the algorithm proposed be feasible in practice within such a modelling framework?

The method proposed, at each time point, draws a model from the posterior distribution over the system dynamics models considered and the estimated optimal allocation strategy is the maximizer over a prespecified class of strategies of the mean outcome under this model. I wonder whether the mean outcome is the right quantity to look at, especially when stochastic transmission models are used. For example, let us consider a homogeneously mixing stochastic SIR model in a closed population. This model has a threshold behaviour which can be described in terms of the distribution of its *final size* which is the number of individuals that ultimately become infected. Fig. 11 demonstrates that the mean final size may mislead the estimation of the treatment. Perhaps, instead of the mean outcome, one possibility is to look at the probability of the epidemic being minor or major.

The authors have done a great job in developing a framework that scales well with the number of locations or units. The (class of) model(s) that they have considered is (are) amenable to direct maximum likelihood estimation and I understand that this plays a key role in being able to explore a large class of strategies. Nevertheless, I believe that further computational challenges still remain to be addressed when the models of the disease dynamics are not that straightforward or quick to fit to the observed data. How to deal with fitting such models of varying complexity, some of which may offer a better fit, is, in my opinion, an important question. Perhaps, fitting a non-parametric model (Kypraios and O'Neill, 2018) might enable us to bypass the need to do model selection in realtime.

Concluding my discussion I congratulate the authors once again for a very interesting paper that brings

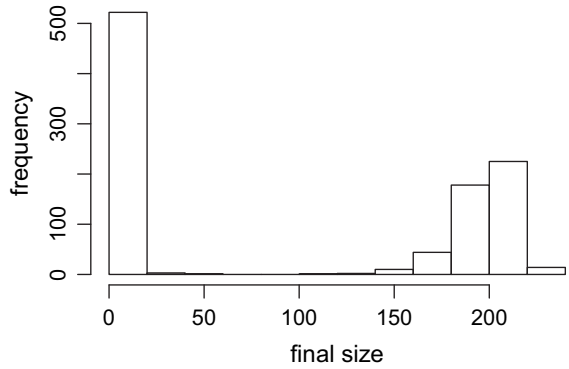

Fig. 11. Final size distribution of a Markov SIR epidemic model in which there are 250 individuals in a closed population, one of which is initially infected and the rest are susceptible; the person-to-person infection rate is 2/250 and each individual remains infectious for some time which is distributed according to an exponential distribution with mean 1 unit

ideas from personalized treatment medicine into developing effective control strategies in spatiotemporal applications.

It therefore gives me great pleasure to propose the vote of thanks.

Daniel Farewell (*Cardiff University*)

I have very much enjoyed reading this paper. I am struck by the complexity of what has been achieved and put in mind of the work of the great radio pioneer Guglielmo Marconi, who brought together many existing components and his own innovations to make a functional system for wireless telegraphy. The authors of this paper have assembled and refined a piece of machinery that, like Marconi's, may save many lives—human, and bat.

I shall attempt to highlight the impressive complexity by way of a little mathematical abstraction. In such problems, I often find it advantageous to employ the notation of stochastic processes, which seem to me ideally suited to both causal reasoning and dynamic control (Dawid and Didelez, 2010). Let $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}, P)$ be a filtered probability space on which we define adapted stochastic processes $X = \{X_t\}$ recording our treatment actions, and $Y = \{Y_t\}$ recording the outcomes of interest. Crucially, we can exert partial or total control over the stochastic process X; by contrast, it is only through X that we can exert any control over Y. The evolution of Y also depends on unknown parameters θ governing the system. Since the authors take a Bayesian approach we understand θ as a random variable defined on the same probability space, and as a variable that covaries with X and Y.

The probability measure P can be decomposed into three main components. For a given intervention strategy σ and at every time t, P_σ specifies the transition measures from the observed information $\mathcal{F}_t$ to increased information $\mathcal{G}_t$ that includes knowledge of our next action X_{t+1} but not its corresponding outcome Y_{t+1}. For a given systems dynamics model θ and at each time t, P_θ governs the transitions from $\mathcal{G}_t$ to $\mathcal{F}_{t+1}$. Finally, Q is the prior distribution on θ.

Let us consider the enormous complexity of the parameter σ, which specifies *at baseline* our treatment strategy *for all time*: at every future time t, how shall we react to the information $\mathcal{F}_t$ in allocating X_{t+1}? Because of its sequential conditioning on available information, the set of possible values of σ is vast. The authors aim to optimize over some subset Σ of this space and compute

$$\arg\max_{\sigma \in \Sigma} E_\sigma\{f(Y)\} \tag{5}$$

for some function f of Y. We index the expectation operator by σ to emphasize its fundamental dependence on the choice of intervention strategy. The daunting complexity of specifying a strategy σ is nonetheless necessary to compute this mean utility; we must fully characterize the distribution under which we are computing expectations. Given this, would the authors agree that an optimal strategy σ is not so much *estimated* as *computed*? Thinking of it as arising sequentially, a realization of X is of course data dependent, but the expectation in expression (5) is not; it is simply a property of P_σ, P_θ and Q.

An ingenious aspect about making expression (5) our objective is the marginalization over θ. The authors write about balancing

> 'exploration of the space of potential allocations with choosing allocations that are estimated to produce high expected utility',

yet this balance is far from evident in expression (5), which focuses entirely on the marginal mean utility. It is a truly remarkable feature of the authors' approach that an application-specific objective combined with a realistic assessment of our uncertainty about θ designs and carries out both experiment and treatment automatically.

A degenerate prior on θ would mean that no stochastic elements remained in the optimal P_σ, comprising at each t a deterministic map from the information $\mathcal{F}_t$ to the known optimal X_t. Thompson sampling exploits this degeneracy and, at time t, samples from current posterior $(\theta \mid \mathcal{F}_t)$. It then computes

$$\hat{\sigma}(\theta) = \arg\max_{\sigma \in \Sigma} E_\sigma\{f(Y) \mid \mathcal{F}_t, \theta\}, \tag{6}$$

although only its most immediate recommendation is needed in practice. As a large part of the paper demonstrates, computing $\hat{\sigma}(\theta)$ is far from straightforward! Fully specified P_σ—the implications of which extend far into the future—are again needed to compute and rank such expectations. Can the authors clarify how this is done? More generally, I would be grateful if they could indicate how we might demonstrate that sampling and optimizing as in equation (6) in fact optimize expression (5), which must depend on the choice of f.

The subset Σ over which we search for optimality can incorporate dynamic constraints: for an $\mathcal{F}_t$-measurable constraint such as the number of treatments currently allowed, we insist that all strategies $\sigma \in \Sigma$ assign no probability to allocations not satisfying this constraint. Implicit in this, though, is some form of model for how such constraints or, more positively, new technologies may be expected to develop in future. Given this, can the authors comment on their suggestion that the class of allocation strategies can adapt to as-yet-unknown treatments becoming available? Does this not require some model for the likelihood of such treatments appearing, and for their hypothetical effectiveness?

G. K Chesterton wrote that 'thanks are the highest form of thought; and that gratitude is happiness doubled by wonder' (Chesterton, 1917). The authors of this paper have given us much to wonder at, and of course a little to wonder about. I am indeed most grateful to them for their wonderful paper, and it is with great happiness that I second the vote of thanks.

The vote of thanks was passed by acclamation.

Bibhas Chakraborty (*National University of Singapore and Duke University, Durham*)

The paper proposes a novel spatiotemporal treatment allocation strategy for controlling an emerging infectious disease by using on-line reinforcement learning. It extends the existing framework of dynamic treatment regimes (DTRs) by incorporating a spatial element into decision making. However, there are some distinctions between the two. First, learning here needs to happen on line, as opposed to mostly off-line, batch mode learning in the case of DTRs (e.g. using 'sequential multiple-assignment randomized trial' clinical trial data). Second, whereas there is scarcity of data for learning at the onset of an epidemic, the number of possible allocations is exponential in the number of locations; this is the inherent 'curse of dimensionality' in the problem. Third, unlike the case of DTRs, independent replications of the data are not available for learning; this resembles the setting of N-of-1 trials. Fourth, there is interference between the spatial locations, violating Rubin's stable unit treatment value assumption for causal inference. Finally, the time horizon can be very long with evolving logistical constraints, resulting in evolving classes of strategies.

The authors have taken a Bayesian approach here, in contrast with the predominantly frequentist DTR literature. The key methodology involves policy search via Thompson sampling. The other methodological innovation is a clever computation of $\arg\max_d C^T(d; \beta; \theta)$ without computing $C^T(d; \beta; \theta)$ for each d within a class, via stochastic approximation; this is critically important for managing the computational burden.

I want to highlight two points that may be missed by some readers. First, even though the approach proposed requires a parametric working model, the focus is solely on finding the optimal strategy, and not on understanding the underlying disease dynamics *per se*. This is precisely where the current approach differs from mathematical modelling. Second, as the authors pointed out, one possible future work is to develop a semiparametric approach based on estimating equations without relying on a parametric

working model. In the case of DTRs (e.g. Q-learning with shared decision rule across time points), we have successfully done something similar (Chakraborty *et al.*, 2016). So I think this research direction is promising.

I would like to pose two questions for the authors' feedback.

(a) The underlying system has been assumed to be time homogeneous. However, in a long horizon problem, the underlying system itself may evolve (e.g. *serotypes* of the disease-causing virus or bacteria). How would you tackle such a problem? Is it simply a matter of computational scalability, or is it more fundamental like the convergence of the algorithm itself?
(b) The methodology proposed works only for completely observed states. The authors mentioned that sampling weights or imputation strategies in conjunction with policy search can handle allocation problems with partially observed states. How different are these from the *partially observable Markov decision process* approaches (Kaelbling *et al.*, 1998) in artificial intelligence?

I conclude by congratulating the authors for a very nice paper.

Robin Henderson (*University of Newcastle, Newcastle-upon-Tyne*)
I congratulate the authors on their fine contribution, which is highly original, impressively broad and thoughtfully practical.

One of the several significant features is the focus on on-line learning concurrent with treatment application: for an emerging epidemic there is just one shot at both finding out what is going on and doing something about it. Do the authors have a view on whether such an updating strategy should also be explored for more standard applications of optimal treatment allocation? I am thinking of personalized medicine, where we can assume independent and identically distributed responses and, so far as I know, the distinction between decision rule determination from training data and subsequent application of policy has always been sharp. Blurring the boundary may be advantageous in some circumstances, e.g. where there are patient-specific parameters, i.e. random effects. A Bayesian approach that starts with population assumptions and then refines posteriors as information for an individual accrues may be beneficial, at least when there are a reasonable number of decision times. Another possibility is to recognize that point estimates obtained from modelling training data are subject to error and to allow continued learning via a stochastic allocation strategy in subsequent practice. Of course this might raise exploration–exploitation ethical issues related to what policy seems to be best right now for a particular patient not necessarliy being better longer term at the group level.

I want to mention also that there are conceptual similarities between the authors' aim of balancing model improvement and optimal treatment choice (item (a) near the end of Section 1) and some of the ideas behind the adaptive model-predictive control procedures that have been developed in the engineering literature, at least for independent data. Control and dynamic treatment researchers often consider very similar types of problem, though from quite different perspectives. More communication between these communities may be beneficial in both directions.

Again, I thank and congratulate the authors for their stimulating paper.

Peter F. Thall (*University of Texas MD Anderson Cancer Center, Houston*)
Laber and his colleagues propose a dynamic strategy for allocating treatments to locations during an epidemic, to limit spread of the disease in space over time, by optimizing a cumulative outcome, such as the number of infected locations or total cost. It is edifying to see a prominent frequentist adopting a Bayesian approach to this challenging problem. Sequentially optimizing functions of accumulated data enables existing dynamic treatment regime machinery to be applied to locations, rather than to individuals, such as patients in a medical setting, except that spatial proximity may introduce interdependence between locations ('interference'). The methodology has several appealing features, including 'Thompson sampling' to avoid becoming stuck at suboptimal strategies, and simulation to validate the methodology. For modelling, in my opinion a Bayesian non-parametric approach should provide the most flexible and robust toolkit.

My main comments pertain to application of the methodology to human epidemics, where the number of deaths and financial cost both should be parts of any pay-off function. As with application of any complex statistical method, numerous general structures must be made specific. The sets of locations and treatment strategies, and data structure to be collected must be determined. Practicalities include the process of data recording, storage, quality control, acquisition and computing the updated optimal allocation rule, all in realtime. Ideally, this could be done via an Internet-based graphical user interface. Since this, and

all other computer software, must be tailored to the application, software development undoubtedly will be a non-trivial, time-consuming process. The methodology certainly will be required to pass political and administrative hurdles involving individuals from multiple regions, states or countries, and these negotiations are likely to involve intensive discussions of pay-offs and resource use. Since multiple decision makers with different utilities must reach a consensus, this in turn will affect key model components. Some combination of elicited expert opinion and historical data may be used to construct priors, and a prior-to-posterior sensitivity analysis will be useful to obtain consensus. This entire process must either

(a) precede the epidemic, which may be unrealistic,
(b) be carried out at some time during a multiyear epidemic, as the authors note in Section 4.

I encourage the authors to think about these issues in the context of applying their methodology to mitigate recurrent annual global influenza epidemics, for which copious data are available. Otherwise, developing this useful methodology seems a little batty.

Marco A. R. Ferreira (*Virginia Tech, Blacksburg*)

I congratulate Dr Laber and his colleagues for their valuable contribution to the area of optimal spatiotemporal treatment allocation for controlling infectious diseases. I have two main concerns that are related to the objective of controlling an infectious disease epidemic and I have a suggestion for possible future research. From the point of view of conservation of the species being affected, I think the main objective should be to ensure the long-term existence of the species. Consequently, my first concern is that I think the long-term existence should have a higher priority than the consistent estimation of an optimal control strategy. Thus, I think we would want to learn as fast as possible about how bad the disease may be, and then we would take action as fast as possible to ensure the long-term survival of the species. With that in mind, observational data (as opposed to stochastic allocation strategies) may provide biased estimates of parameters of disease dynamics but may be sufficiently good for making timely decisions on how to control the disease. An alternative way to look at this first concern would be that the amount of time needed to make a decision (including time for data collection) should be part of the total utility function. My second concern is related to the discounted total utility that depends on a discount parameter $\gamma \in (0, 1)$; this utility function discounts the future, and as such would weigh more heavily the near future than the long-term existence of the species. The authors mention that their methodology can be extended to handle a cumulative utility function such as $\lim_{T\to\infty} \Sigma_{t=1}^{T} u(\mathbf{Y}^T)$ which I think would be more in line with ensuring the long-term existence of the species. It would be helpful if the authors could elaborate on how to accomplish such an extension. Finally, the problem of finding an optimal allocation strategy belongs to the class of optimal Bayesian design problems and, as such, requires a maximization of a multi-dimensional integral. In light of this, a possible avenue for future research may be the extension of currently available stochastic algorithms that perform joint maximization and integration such as the inhomogeneous Markov chain algorithm (Müller *et al.*, 2004) and the inhomogeneous evolutionary Markov chain Monte Carlo algorithm (Ferreira and Sanyal, 2014; Ferreira, 2015) to the context of optimal allocation strategies for controlling infectious diseases.

Chengchun Shi, Rui Song and Wenbin Lu (*North Carolina State University, Raleigh*)

We congratulate the authors for their very thoughtful paper on on-line decision making for infectious diseases. In addition to Thompson sampling, many other algorithms such as the ϵ-greedy algorithm (Sutton and Barto, 1998) and the upper confidence bound type algorithm (Agrawal, 1995) can be applied to on-line decision making as well. Taking the ϵ-greedy algorithm as an example, at the jth treatment stage, given the current parameter estimators $\tilde{\beta}^j$ and $\tilde{\theta}^j$ and some $0 < \epsilon_j < 1$, we may set $\pi^j = \arg\max_{d\in\mathcal{D}} C^T(d; \tilde{\beta}^j, \tilde{\theta}^j)$ with probability $1 - \epsilon_j$, and randomly sample π^j from $\mathcal{D}$ with probability ϵ_j. It might be interesting to compare Thompson sampling with these algorithms.

In step 6 of algorithm 1, the authors use policy search to compute $\hat{\pi}^{j+1}$ by maximizing the estimated value function among a class of treatment allocation strategies indexed by a finite dimensional vector. Policy search (Zhang *et al.*, 2012, 2013) is a recent popular method for estimating the optimal treatment rule. We wonder whether classical reinforcement learning methods such as Q-learning (Watkins and Dayan, 1992) could be used to estimate the optimal allocation strategy. From a practical perspective, it will be helpful if the authors could compare policy search with Q-learning under the set-up of controlling emerging infectious diseases.

Under assumptions 1–3, the authors show in equation (1) that the average discounted utility function

under a given policy π can be expressed by using the data-generating model. We are curious whether the authors can present more details on this equation. Why does it hold under the given assumptions?

In algorithm 2, the authors use a stochastic approximation algorithm to approximate $\arg\max_d C^T(d;\beta,\theta)$. The algorithm relies on a sequence of non-negative step sizes $\{\alpha_j\}_{j\geqslant 1}$ and the authors comment that convergence guarantees require $\Sigma_j \alpha_j = \infty$. However, in the implementation, the authors use $\alpha_j = \tau/(\rho + j)^{1.25}$. We note that $\Sigma_j \tau/(\rho + j)^{1.25} < \infty$ for any $\tau, \rho > 0$. As a result, the algorithm may not converge under such choices of $\{\alpha_j\}_j$. Is it possible to use $\alpha_j = \tau/(\rho + j)$ to improve the convergence of the algorithm?

M. Elizabeth Halloran (*Fred Hutchinson Cancer Research Center and University of Washington, Seattle*) **and Michael G. Hudgens** (*University of North Carolina at Chapel Hill*)

Laber and his colleagues are to be congratulated for developing this fine framework for realtime data-driven decisions for managing emerging infectious diseases. Motivated by the white nose syndrome epidemic in bats, they develop methods for determining optimal on-line treatment allocation in space and time. The methods allow for interference, i.e. treatment of a location can affect the potential outcomes in another location (Cox, 1958).

Historically, interference can be placed in the context of what Ross (1916) called dependent happenings. In his 'theory of happenings', Ross differentiated independent from dependent happenings. Dependent happenings are those in which the frequency depends on the number already affected. To this class belong

> 'infectious diseases, membership of societies and sects with propagandas, trade-unions, political parties, etc., due to propagation from within, that is, individual to individual'

(Ross (1916), page 211).

In this paper the propagation of infection is transmission from one spatial location to another. Interference and spillover effects are governed by a gravity model describing the probability of spread of white nose syndrome between cave-bearing counties. A challenge is that the complexity of determining the optimal treatment allocation grows quite large with the number of locations because the number of potential treatment allocations grows exponentially. The situation is similar when studying effects of vaccination of individuals in populations in the presence of interference. Laber and his colleagues use a clever combination of theory and computational methods to reduce the complexity by constructing a class of flexible strategies that are highly scalable. The cumulative expected utility at a time t, e.g. the proportion of locations infected at time t, is compared under different strategies, where their proposed optimal strategy slows spread better than two other rather rote strategies they call proximal and myopic.

Intervention effects are defined within the gravity model describing the local dynamics. Treatment can reduce the transmission probability both from a treated infected location (reduced infectiousness) and to a treated uninfected location (reduced susceptibility). However, treatment does not cure an infected location but only slows a process wherein every location becomes infected eventually, which is a depressing prospect for the bats. A simple extension would include that treatment could cure an infected site, allowing for elimination. We agree with the authors that one important avenue of future research will be developing methods that are robust to misspecification of the system dynamics model, e.g. perhaps by using Bayesian model averaging over multiple system dynamics models.

Daniel J. Lizotte (*University of Western Ontario, London*)

This work is an excellent example of applied statistics as the 'art of the possible'. Starting from well-established formalisms of sequential decision making, the authors develop practical methodology, and along the way they contribute to our understanding of how to bring these formalisms to practical application. Their work is also an important reminder, perhaps more salient to the computer science community than the statistics community, that not all important problems are 'big data' problems, and not all 'small data' problems have trivial solutions. I would like to discuss an additional application area that may benefit from this work.

There is significant interest in health research on the topic of multimorbidity, defined as the presence of multiple chronic diseases in the same individual. Such diseases often share a complex web of causes, and their associated treatments often interact with each other as they shape an individual's future health. However, thinking about a 'joint treatment plan' for several diseases simultaneously in many areas of medicine is still cutting-edge research and often has little data-driven evidence to support its practice directly. The formalism developed by the authors—particularly the assumption of independent action control but dependent effects across units in the future—is well suited to this problem. At the same time, the multimorbidity treatment problem is qualitatively a little different. We might think of different diseases

as 'regions', each with a *repertoire* of actions that may be complex in and of itself. However, the number of diseases or regions would typically be small (fewer than 10), and the propagation of treatment effects across and among them could be very complex. This would require additional methodology to capture effectively. There are important 'art of the possible' considerations in this domain that must be addressed as well. The output of any system for the multimorbidity treatment problem can only be decision *support* rather than decision making, which in part means that uncertainty over different courses of action must be exposed in a way that end-users understand. This might be accomplished by using the authors' framework and somehow conveying the posterior distribution over optimal allocations. Furthermore, any recommended treatment may or may not be followed—an issue alluded to by the authors in their discussion of a 'compliance model'. I am interested to know the authors' opinions on their work's relevance to other such sequential decision-making problems with complex action spaces.

The following contributions were received in writing after the meeting.

Anna L. Choi (*Chinese University of Hong Kong in Shenzhen and Shenzhen Research Institute of Big Data*) **and Tze Leung Lai** (*Stanford University*)

Section 2 of this interesting paper gives a comprehensive review of white nose syndrome in north American bats. Section 3 formulates the authors' proposal to control the spread of white nose syndrome by using optimal treatment allocation. For each time point and each location, the decision is to 'apply a treatment or to do nothing'. It considers only one treatment with unknown treatment effects. However, there are already many treatments that are yet to be tested (Cornelison *et al.*, 2014; Hoyt *et al.*, 2015), and the (economic) cost of the treatment(s) should also be considered. In this connection, a relatively economical and ease-to-use new treatment has just emerged at the beginning of 2018. *Science Daily*, January 2nd, 2018, announced that

> 'scientists with the USDA Forest Service and the University of New Hampshire have found what may be an Achilles' heel in the fungus that causes white-nose syndrome: UV-light',

citing a study by Palmer *et al.* (2018). This adds new challenges to the optimal treatment allocation problem as new treatments can enter the treatment pool during the course of the study.

We next comment on the more general problem of optimal allocation of a (single) treatment 'over a countably infinite set of treatment periods and a finite number of locations'. The solution is described as 'estimating an optimal allocation strategy' in Section 4, following the framework of optimal dynamic regimes in the statistics literature. In control engineering, there is a counterpart called *stochastic adaptive control*; in computer science, the counterpart is *reinforcement learning*. The fascinating sequential treatment allocation problem over a dynamic network system, with covariate and outcome information from relevant nodes of the network, considered in this paper also arises in many other interdisciplinary applications on which we are writing a monograph (Choi *et al.*, 2019). In this connection, we want to point out that the authors' remarks near the end of Section 7 fall under the framework of *contextual bandits*, for which it is shown that Thompson sampling is not optimal because it is myopic whereas a much simpler ϵ-greedy strategy can be shown to be asymptotically optimal.

Dean Eckles (*Massachusetts Institute of Technology, Cambridge*) **and Maurits Kaptein** (*Tilburg University*)

Laber and his colleagues impressively model intervening to prevent epidemic spread as a multiarmed bandit problem, using Thompson sampling to add exploration to the choice of sites to treat. We wish to highlight how

(a) the method could be made more robust by use of a bootstrap and
(b) how less exploration will often be preferable with such a short horizon.

The authors approximate Thompson sampling by using a plug-in estimator of the sampling distribution of the maximum likelihood estimates (Section 5.1). One could further depart from trying to approximate the posterior of the posited parametric model by using a non-parametric bootstrap distribution (see Newton and Raftery (1994) and Efron (2012)). Bootstrap Thompson sampling (Eckles and Kaptein, 2014) samples from replicates formed by an on-line bootstrap, which can be easily parallelized or distributed, can account for dependent observations and is more robust to common forms of model misspecification. This method and related greedier variations have been used in multiple application areas and compared favourably with other methods (e.g. Agarwal *et al.* (2014), Osband *et al.* (2016), Lu and Van Roy (2017) and Bietti *et al.*

(2018)). In the context of the present application, this could enable the authors to avoid Markov chain Monte Carlo sampling altogether while being robust to, for example, heteroscedastic errors or within-region dependence.

We have already mentioned the popularity of greedier variations of bootstrap Thompson sampling, More generally, recent work has demonstrated that methods that explore less (i.e. are greedier) perform extremely well in contextual multiarmed bandit problems (Chapelle and Li, 2011; May *et al.*, 2012; Bastani *et al.*, 2018; Bietti *et al.*, 2018) with common characteristics. First, in many applications the horizon T is small; in the present application, actions are taken for eight periods only. With such a small horizon, it is not obvious that substantial exploration is preferable—at least for minimizing regret. In contrast, some other recent results in favour of exploration-free methods based on the distribution of context (e.g. Bastani *et al.* (2018)) do not obviously apply to the present setting with dependence due to contagion and with combinatorial actions. Nonetheless, we expect that a policy that is more greedy than the presented implementation of Thompson sampling would lead to an even smaller proportion of infections.

The best reason for introducing exploration may not be minimizing regret up to the horizon. Rather, introducing relatively low cost exploration might be motivated by the desire to make reuse of collected data possible for other purposes. In standard contextual multiarmed bandit and dynamic treatment regime settings, reuse of such data for evaluation of other polices is routine (Murphy *et al.*, 2001; Dudík *et al.*, 2014; Agarwal *et al.*, 2016). It would be interesting to see how readily data from a setting with dependence induced by contagion could be reused.

Seongho Kim (*Wayne State University, Detroit*) **and Weng Kee Wong** (*University of California at Los Angeles*)

We congratulate the authors on their detailed and interesting work in this important area of research.

Assumptions 1 and 2 would seem to imply that $\mathbf{A}^t \perp \mathbf{W}^t | \mathbf{H}^t$ for all $t \in T$ and so $\mathbf{A}^t$ is independent of $\mathbf{Y}^t$, $\mathbf{S}^{t+1}$ given $\mathbf{H}^t$ for all $t \in T$. Further, by the Markov homogeneity assumption,

(a) the observed data $\mathbf{S}^t$ and allocation $\mathbf{A}^t$ at time t are sufficient to predict the outcome at time t, $\mathbf{Y}^t$, and
(b) the data at time t, $\mathbf{S}^t$, depend only on the observed data, $\mathbf{S}^{t-1}$, and the allocation $\mathbf{A}^{t-1}$ at time $t-1$.

Fig. 12 depicts this time-dependent homogeneous Markov system but suggests that the conditional independence assumption between $\mathbf{A}^t$ and $(\mathbf{Y}^t, \mathbf{S}^{t+1})$ given $\mathbf{H}^t$ may be questionable because of the edges with asterisks and, if so, will appear to contradict assumptions 1 and 2. To avoid this discordance, it seems that either assumption 1 may be eliminated or needs to be modified to $\mathbf{Y}^{*t} \perp \mathbf{S}^{*t+1} | (\mathbf{A}^t, \mathbf{H}^t)$ for all $t \in T$.

At March 14th, 2018, white nose syndrome (WNS) had spread to Washington, but not to Florida (`https://www.whitenosesyndrome.org/sites/default/files/wnsspreadmap_3_14_2018.jpg`). Florida, near to the contaminated regions, has not reported WNS yet, whereas Washington has been infected without neighbouring infected regions. Can the spatial gravity model incorporate such information in equation (2) and, if so, how does the model proposed update the parameters by using a Bayesian framework with the latest WNS data? These are likely to be challenging tasks. It is also possible that symptoms from bats in Florida are temporary because of the relatively high average temperatures

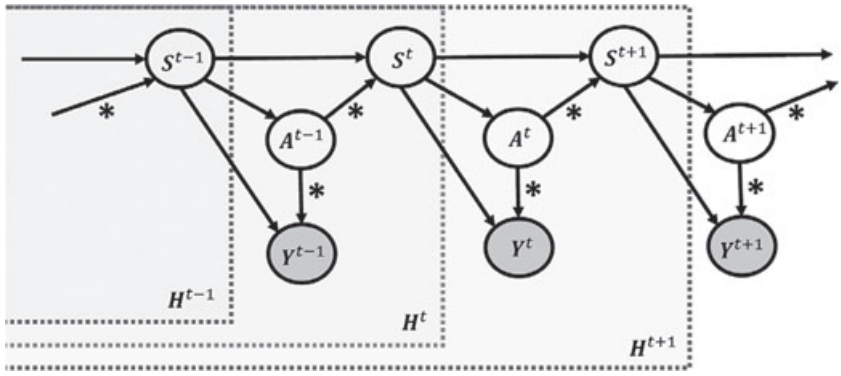

Fig. 12. Graphical representation of the dependence between $\mathbf{S}^t, \mathbf{A}^t, \mathbf{Y}^t$ and $\mathbf{H}^t$ where $t \geqslant 1$, $\mathbf{H}^t = (\mathbf{S}^1, \mathbf{A}^1, \mathbf{Y}^1, \ldots, \mathbf{S}^{t-1}, \mathbf{A}^{t-1}, \mathbf{Y}^{t-1}, \mathbf{S}^t)$ and $\mathbf{H}^1 = \mathbf{S}^1$

(Verant *et al.*, 2012) and so WNS in bats cannot be confirmed without direct microscopy and culture analyses. If so, Florida would have neither uninfected nor infected areas. Having models with three statuses for each outcome **Y**, uninfected, symptomatic infected and asymptomatic infected status, may be helpful. Another possibility is to extend the model to have a parallel pathogen model with New York and Washington as parallel pathogens, and/or a model with multiple transmission pathways (Tien and Earn, 2010).

There are a few covariates in equation (2) and having more covariates might improve the model predictive ability. With the additional covariates, the number of parameters will increase, which may result in singularity problems. We would like to know whether the authors have tried using lasso-type penalization in equation (2) and their thoughts on global optimization and identifiability, which are two critical issues working with high dimensional non-linear models.

Johan Koskinen (*University of Manchester*)

This very thorough and welcome contribution demonstrates the use of simulation in modern inference and engages with several academic traditions, one of which, social network analysis, I would like to add some reflections on. Spread on networks has a long history in social network analysis (Coleman *et al.*, 1957) and, particularly in health sciences, much recent attention has been paid to interventions (Valente, 2012; Morris, 2004). Estimating these processes has prompted the development of statistical models (such as Greenan (2015)) and simulation routines (e.g. Jenness *et al*. (2016)). Rolls *et al.* (2012) investigated the behaviour of hepatitis C transmission on an empirical network (Aitken *et al.*, 2008) and demonstrated effects of different types of network topologies on disease outcomes. Information on these disease relevant networks is by necessity patchy but can be learned through sampling techniques such as in Rolls *et al.* (2013a) or Handcock and Gile (2010). In fact, key features can to some extent even be estimated from respondent interviews (Krivitsky and Morris, 2017). Many of these methods rely on exponential random-graph models (Lusher *et al.*, 2013), a class of log-linear models that model interactions between links that correspond to features such as the prevalence of hubs and the tendency towards clustering of ties. Fig. 13 illustrates potential consequences of such features in a 'realistic' network (Fig. 13(b)) relative to an 'unrealistic' network (Fig. 13(a)) with the same number of links. The local clustering of network (b) means that links, rather than carrying the disease further from the seed node (black), out into the network, are used up linking to nodes that are already connected. Real life networks—both human and animal (even bats; see for example Willis and Brigham (2004)—typically demonstrate high clustering (not merely artefacts of geography; Daraganova *et al.* (2012)) and relatively short path lengths. The example network N3 is meant to have these features (Robins *et al.*, 2005) but to the naked eye appears to have rather long pathways. In fact the results, pairwise for S1 and N1, S2 and N2, and S3 and N3, seem to reflect the fact that the network topologies are more stringent forms of linkage than their spatial equivalents, resulting in similar behaviour but with lower uncertainty. How would the spread and interventions be affected by the more distinct network features? Obviously, these are difficult and complex questions (Rolls *et al.*, 2013b) but the proposed allocation scheme is also very flexible.

Michael T. Lawson, Hunyong Cho, Arkopal Choudhury, Yifan Cui, Xiaotong Jiang, Teeranan Pokaprakarn and Michael R. Kosorok (*University of North Carolina at Chapel Hill*)

We present brief explorations of three topics in the paper that we believe may lead to interesting future research.

The authors' use of the rank-based priority function $\mathcal{R}$ heightens the importance of the constraint c^t in estimating π. Although this paper assumes c^t is given, in practice c^t may be more flexible, and different

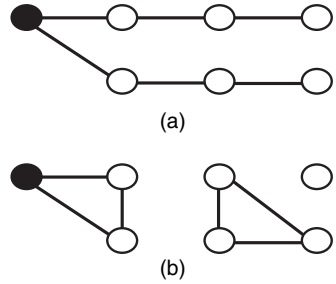

Fig. 13. Two schematic network structures with the same number of nodes and ties but different numbers of closed triads

specifications of c^t may greatly alter the course of disease. We think this resource allocation problem—determining how much of a pool of resources to spend at each time point—provides an important and challenging direction for future research in the context of infectious diseases. One possible approach is to include the number of caves to treat at each time point in the action space and to incorporate the cost of the actions in the utility function.

Approximate Bayesian computation (ABC) based on sequential Monte Carlo (SMC) sampling (Rubin, 1984; Toni *et al*., 2009) may prove a faster alternative for estimating the posterior distribution of (β, θ). ABC SMC sampling draws large groups of candidate predictors simultaneously to avoid local minima and to speed up computation. ABC SMC sampling offers several attractive features of Bayesian analysis, including the ability to quantify the uncertainty in predictions (McKinley *et al*., 2018) and to account for missing data with data augmentation methods that add potentially unobserved states to the parameter space (O'Neill and Roberts, 1999), or faster alternatives to data augmentation such as modified 'poor man's data augmentation' algorithms (Sweeting and Kharroubi, 2005). Sufficient statistics can be of key importance in ABC methods, and finding the correct such statistic in spatial epidemic models may provide a fruitful avenue for future research.

In white nose syndrome, measurements are costly but prone to error, and individual observations can have a substantial downstream effect. As such, although measurement accuracy is of paramount concern, a 'measure until sure' approach is likely to be infeasible. A more realistic approach is targeted remeasurement: only remeasure in the locations where it is most crucial to do so. We propose a measurement influence statistic, whose rationale is akin to that of predictive inference methods (Geisser, 2017). Adopting the notation from the paper we calculate the influence for location l at time t as follows.

(a) Perturb Y_l^t to the opposite disease status.
(b) Calculate $R_{(l)}(s^t, a^a; \eta)$, the vector of priority scores when l is perturbed.
(c) Calculate $\Delta_l^t = \delta\{R(s^t, a^t), R_{(l)}(s^t, a^t)\}$, where δ is an appropriate distance metric.

Observations with the largest Δ_l^t disrupt the priority scores the most when they are mismeasured, assuming that all other locations are measured correctly, meaning that they are important targets to remeasure.

Jorge Mateu (*University Jaume I, Castellón*)

The authors are to be congratulated on a valuable contribution and thought-provoking paper on optimal allocation strategies for control of emerging infectious deseases. I shall focus my discussion on several aspects of dynamical systems on networks, as these are one of the main supports for the evolution of epidemics on individuals. The authors propose a statistical framework to study sequential treatment allocation identifying several major computational and theoretical challenges.

Recently a new algorithm for sampling posteriors of unnormalized probability densities, called 'ABC shadow', was proposed in Stoica *et al*. (2017). In the optimal treatment allocation within the context of dynamical systems, we have a probability density $p(y|q)$ which is strictly positive and continuously differentiable with respect to q. Stoica *et al*. (2017) introduced a global optimization procedure based on the ABC shadow simulation dynamics. They proposed a simulated annealing method to maximize probability densities with unknown normalizing constants that are not available in analytic closed forms. Thus, special strategies are required to sample from the posterior distribution of such probabilities. A combination of simulated annealing with the ABC shadow dynamics is easily adapted to dynamical systems on networks, reducing both computational complexity and time.

The authors introduce an allocation strategy combining simulation optimization with Thompson sampling. Their method could be improved by using the two-step method in Khavarzadeh *et al*. (2018) that performs spatiotemporal balanced sampling in a design-based approach making use of a three-dimensional Voronoi tessellation. This method fits particularly well under the presence of spatiotemporal trends and/or anisotropic effects in the variable of interest: cases often found in dynamical systems.

One prominent topic in the analysis of complex structures is the relationships in network structures, where the object of interest lies in the investigation of structures between different entities of the network. Eckardt and Mateu (2018) consider an alternative formalism that allows taking undirected, directed or partially directed graph structures as well as temporal dynamics into consideration. This alternative approach highlights the possibility of achieving several different graph-based intensity formulations and related statistics for network structures. Eckardt and Mateu (2018) assume that a point pattern appears randomly between pairs of georeferenced nodes whose location within a planar region is treated as fixed. Graphical models appear as a natural choice to explore conditional interrelationships between spatial observations, and we claim they can play a big role in this context of dynamical systems on networks.

James McGree and Kerrie Mengersen (*Queensland University of Technology, Brisbane*)

Effective control measures are important for minimizing the effect that infectious diseases can have on human and animal welfare, and we congratulate the authors on a very interesting paper. We also stress, as the authors quite rightly point out, the importance of understanding the dynamics of epidemics for developing effective control measures. Unfortunately, through adopting utility functions based only on, for example, the number of infected individuals or locations, little provision is made for learning about the process which is being controlled, limiting the information that can be obtained about the epidemic. This can significantly hamper the development of optimal control strategies, and the future development of detection and prevention measures, all of which are important components of managing infectious diseases.

Decision theoretic approaches in Bayesian design provide the methodology to control and learn about an epidemic simultaneously; see Ryan *et al*. (2016) for a recent review. In the context of controlling infectious diseases, this could include learning about

(a) how the disease spreads spatially and temporally,
(b) what subpopulations exist within the population and how individuals transition between them, and
(c) important parameters such as transmission probabilities and how these may vary depending on individual or location information.

Learning about (b), for example, could provide a mechanistic understanding of the disease which can be exploited in the development of more targeted control strategies. This should significantly reduce the uncertainty within the set of competing models, leading to more effective treatment allocations. Such understanding would enable health authorities to be more proactive rather than reactive in controlling infectious diseases.

Considering such a decision theoretic approach would require extending the author's utility function to encapsulate information-based utilities appropriately such as mutual information for parameter estimation and model discrimination (Shannon, 1948; Box and Hill, 1967; Borth, 1975). This will result in a utility function which is no longer a summary of prior predictive data but rather a functional of the posterior distribution resulting from prior predictive data. This imposes considerable computational challenges which could be addressed in the future through the development of fast and/or approximate inference methods, and may also require the further development of optimization algorithms that can handle noisy and expensive utility functions. Such challenges reside at the intersection of optimal control and Bayesian design, and would be an interesting area to pursue into the future with the developments proposed by the authors of this paper.

Erica E. M. Moodie and David A. Stephens (*McGill University, Montreal*)

We read with great interest Laber and his colleagues' on-line dynamic treatment allocation strategy. The majority of the statistical literature on dynamic decision making has focused on binary treatment decisions with the observed units independent and many relative to the number of parameters to be estimated. Although some methods are amenable to continuous doses (Murphy, 2005; Robins, 2004; Rich *et al*., 2016; Chen *et al*., 2016) or multiple treatment options (Tao and Wang, 2016; Tao *et al*., 2018; Zhou *et al*., 2018), the information relative to the complexity of the problem is still typically large.

The approach that Laber and his colleagues have taken allows for new treatments to be incorporated in the on-line procedure as they become available and allows for constraints to be imposed on the treatment process. This approach has enormous potential implications for the control of illness in human populations. Consider, for example, modelling treatment to contain the spread of nosocomial infections where network edges could be defined by, for example, patients seen by the same physician (room-to-room spread), or by ward centroid. To tackle this problem, further statistical refinements are needed: space may be three dimensional if floor-to-floor transmission is possible, or it may have discontinuities if it is deemed *a priori* to be unlikely. Further, there will be a need to decide not only which patients to treat, but with which drugs (e.g. narrow or wide spectrum antibiotics). Drawing on related literature such as the dose finding trials of Lee *et al*. (2015) may prove a good starting point.

We congratulate Laber and colleagues on their innovation addition to the dynamic decision-making methodology, and we encourage researchers in the field to consider the additional statistical challenges that we have noted. We are approaching an era where algorithms are routinely used to decide patients' treatment, though most are static in time (e.g. the software platforms used to determine warfarin dosing, or a recently developed suicide risk predictor (Simon *et al*., 2018)). As on-line algorithms move into clinical practice, many practical details must first be in place: consistent electronic recording of information at

clinical encounters and meticulous and explicit documentation of *which* treatment rule is being used at a given point in time so that—in the event of litigation or for research—the particular instance of the learned, on-line algorithm being used is known. As statisticians, we cannot hope to solve all of these problems, but continued dialogue with end-users is more important than ever as the statistical study of precision medicine is moving ever more rapidly into real world implementation.

The **authors** replied later, in writing, as follows.

We are immensely grateful to the Royal Statistical Society for the opportunity to present and discuss this work. We are also indebted to all those who provided feedback on our work including the reviewers and the discussants. This feedback has deepened our understanding and raised some interesting open problems. In this rejoinder, we discuss some of these problems which we have organized by topic rather than by individual discussant.

The exploration–exploitation trade-off

Several discussants suggested ϵ-greedy or upper confidence bound (UCB) sampling as alternative approaches to managing the so-called exploration–exploitation trade-off (Chakraborty; Choi and Lai; Shi and his colleagues). These approaches have been used to great success in a wide variety of application domains (see Kaelbling *et al*. (1996), Si *et al*. (2004), Wiering and van Otterlo (2012) and references therein); however, in the context of controlling an emerging infectious disease, Thompson sampling (TS) is appealing for the following reasons.

(a) *TS avoids pure exploration steps*: at the time of outbreak, when the disease is confined to a small subset of a large spatial domain, effective allocations are often tightly clustered around the infected locations. However, the set of such allocations can be a vanishingly small fraction of the $O(2^L)$ set of all possible allocations. In these settings, drawing an allocation uniformly at random can lead to poor disease control with high probability while yielding little information to improve the underlying model Fig. 14 shows the schematic diagram for an infectious disease spreading across a 30×30 grid of locations. At onset, there are two infected locations denoted by the filled cells in the lower left-hand side of the grid. Under a diffusive disease model where it is very unlikely for the disease to spread more than, say, one or two squares during the next time step, and assuming that treatment is effective only while it is being applied, one should limit allocations to those near the infection, e.g. inside the 5×5 square outlined, as treatments outside this region are unlikely to slow the spread of the disease or to generate information about the underlying treatment effect. If resources allow for the treatment of 10 locations, then an allocation sampled uniformly at random (as in an exploration step of an ϵ-greedy algorithm) will treat five or more locations within this square with probability approximately 2.5×10^{-6}.

In addition, the prospect of sampling treatment locations independently from available informa-

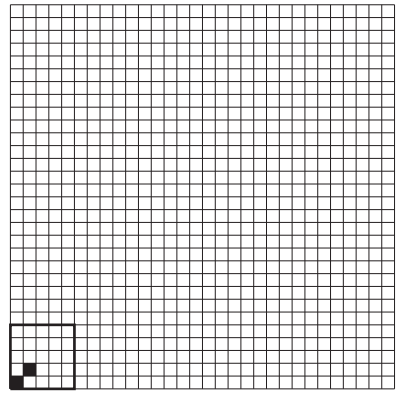

Fig. 14. Schematic diagram for an infectious disease spreading across a 30×30 grid: at onset, the disease has infected the two filled cells in the lower left-hand corner; under the assumption that at most 10 locations can be treated at each time step, the probability that a uniformly sampled allocation will have five or more treatments inside the 5×5 square outlined is approximately 2.5×10^{-6}

tion about the underlying disease dynamics may not be acceptable to (nor considered to be ethical by) policy makers. However, in the context of an on-going human epidemic, even unequal randomizations like those used in TS may also be unpalatable (Thall). Below we discuss tuning the degree of randomization in TS.

(b) *TS leverages scientific theory*: data are scarce at the time of outbreak and the states and allocations are high dimensional. Thus, it is critical to impose sufficient structure on the disease dynamics model and the class of allocation strategies to make the problem tractable. As noted above, ϵ-greedy algorithms ignore both domain knowledge and any accumulated data during exploration steps by sampling allocations uniformly at random. UCB sampling makes use of the underlying postulated model in computing the confidence bounds as follows. As in the main paper let $C^T(d;\beta,\theta)$ denote the (truncated) utility associated with strategy $d \in \mathcal{D}$ under a generative model indexed by (β,θ) and $\{\nu^t\}_{t\geqslant 1}$ denote a non-increasing sequence in (0, 1). Define $\mathfrak{D}^t_{\nu^t}(d)$ to be the $100(1-\nu^t)$ percentile of the posterior distribution of $C^{\mathrm{T}}(d;\beta,\theta)$ given the data accumulated at time $t \geqslant 1$; then UCB sampling selects the allocation strategy

$$\breve{d}^t = \arg\max_{d\in\mathcal{D}} \mathfrak{D}^t_{\nu^t}(d), \tag{7}$$

and subsequently chooses allocation $A^t = \breve{d}^t(\mathbf{S}^t)$. Because the bound $\mathfrak{D}^t_{\nu^t}(d)$ captures both the estimated mean utility under d and the uncertainty of this estimator, it thereby manages the exploration–exploitation trade-off (Lai and Robbins, 1985; Audibert *et al.*, 2009). The primary challenge associated with UCB algorithms in the context of emerging infectious diseases is the computational burden that is associated with computing equation (7). A naive approach to estimating $\mathfrak{D}^t_{\nu^t}(d)$ is

(i) to sample a large number of draws from the posterior over the parameters (β,θ),
(ii) for each sampled parameter value to use Monte Carlo sampling to estimate the mean discounted utility under d and
(iii) to estimate the $100(1-\nu^t)$ percentile of the posterior distribution mean discounted utility.

Steps (ii) and (iii) would need to be repeated at each iteration of a stochastic optimization algorithm which could be costly in large problems. However, it may be possible to use parallelization, emulation or other approaches to reduce the computational cost.

(c) *TS is a general and highly extensible framework*: the Bayesian underpinnings of TS can gracefully handle accumulating spatiotemporal data. Although we considered susceptible–infected models in our applications, as several discussants note (Kypraios; Halloran and Hudgens), these models can be generalized to susceptible–infected–removed or even more general models (e.g. non-parametric Bayesian models as suggested by Thall). Furthermore, although we used the discounted marginal mean utility as our objective one could instead use the average outcome over a finite or indefinite time horizon or more generally any real-valued function of the accumulated history (Kypraios; Ferreira).

In terms of implementation, that TS requires only a single draw from the posterior at each time point provides some computational gain relative to UCB or other Bayesian reinforcement learning methods (Ghavamzadeh *et al.*, 2015). This computational gain comes at the expense of potentially worse performance relative to a less myopic (fully) Bayesian procedure that attempts to account for how the posterior will change in the future as the result of current actions (Choi and Lai; Guez *et al.* (2014)). At the extreme, as Farewell suggests, one can attempt to optimize over the entire space of possible learning algorithms given only the prior and class of allocation strategies, though it is not clear how to optimize over this space.

The preceding advantages were among the motivating factors for developing and applying a spatiotemporal variant of TS to the management of white nose syndrome. We envisioned a continually adapting allocation strategy that could be used indefinitely as a decision support tool for policy makers. However, Eckles and Kaptein note that, in some settings, it may be more realistic to plan for a finite time horizon in which case TS may be exploring too much. This is easily seen in the extreme case of a single decision point wherein exploration provides no benefit. Eckles and Kaptein also suggest the bootstrap as an approach to managing on-line exploration (Eckles and Kaptein, 2014). We discuss one approach to balancing the amount of exploration by using a multiplier bootstrap (Praestgaard, 1990; Van der Vaart and Wellner, 1996; Kosorok, 2008). The score function for (β,θ) given data up to time t is

$$\sigma^t(\beta,\theta) = \sum_{v=1}^{t} \nabla_{\beta,\theta} \log\{ f(\mathbf{Y}^v|\mathbf{S}^v, A^v;\beta) g(\mathbf{S}^v|\mathbf{S}^{v-1}, A^{v-1};\theta)\};$$

thus, the maximum likelihood estimators $(\hat{\beta}^t, \hat{\theta}^t)$ solve $\sigma^t(\beta, \theta) = 0$. Let $\mathbf{M}^t = M_1^t, \ldots, M_t^t$ denote non-negative independent and identically distributed (IID) random variables with mean and variance 1, and define

$$\sigma^t_{M^t}(\beta, \theta) = \sum_{v=1}^{t} M_v^t \nabla_{\beta,\theta} \log\{f(\mathbf{Y}^v|\mathbf{S}^v, A^v; \beta) g(\mathbf{S}^v|\mathbf{S}^{v-1}, A^{v-1}; \theta)\};$$

let $(\hat{\beta}^t_{M^t}, \hat{\theta}^t_{M^t})$ solve $\sigma^t_{M^t}(\beta, \theta) = 0$; then the conditional distribution of $(\hat{\beta}^t_{M^t}, \hat{\theta}^t_{M^t})$ given the observed data is a multiplier bootstrap approximation to the sampling distribution of $(\hat{\beta}^t, \hat{\theta}^t)$. Using the estimated sampling distribution as a surrogate for the posterior, we obtain the following bootstrap analogue of TS:

(a) sample $\mathbf{M}^t$;
(b) solve $\sigma^t_{M^t}(\beta, \theta) = 0$ to obtain $(\hat{\beta}^t_{M^t}, \hat{\theta}^t_{M^t})$;
(c) set $d^t = \arg\max_{d \in \mathcal{D}} C^T(d; \hat{\beta}_{M^t}, \hat{\theta}^t_{M^t})$.

To control the exploration–exploitation trade-off one might consider adjusting the distribution of the bootstrap weights, $\mathbf{M}^t$; for example, we could select these to follow a gamma distribution with mean 1 and variance $\tau \geqslant 0$. Setting $\tau = 1$ corresponds to the Bayesian bootstrap whereas setting $\tau = 0$ corresponds to greedy allocation selection (i.e. no exploration). Thus, values of $\tau \in [0, 1]$ represent a spectrum of TS-like algorithms with varying degrees of exploration. One can imagine adaptive variants of such algorithms wherein the variance decreases to 0 as t approaches a finite time horizon. Computing such a sequence as a function of the prior, the class of strategies and the time horizon would be a special case of the framework suggested by Farewell. We think that the study of such algorithms is an exciting direction for future research.

Methodological extensions

Several discussants mentioned ways in which the modelling framework proposed might be extended (Kypraios; Chakraborty; Thall; Halloran and Hudgens). As noted above, a benefit of the underlying Bayesian framework is the ability to deal with accumulating data that may be subject to missingness (Little and Rubin, 2014), measurement error (Carroll *et al.*, 2006), partial observability (Poupart and Vlassis, 2008; Ross *et al.*, 2008) and non-stationarity (West and Harrison, 2006); Lawson and his colleagues proposed an interesting alternative to deal with measurement error that might be particularly effective in the context of white nose syndrome. However, our proposed methodology and the foregoing extensions suppose that the postulated models for the underlying data-generating process are correctly specified. Thus, another natural question is to what extent the methodology proposed is sensitive to misspecification (Halloran and Hudgens) and whether semiparametric methods, e.g. Q-learning (Murphy, 2005; Schulte *et al.*, 2014), might provide a more robust alternative (Chakraborty). Here we describe a semiparametric variant of TS based on the Q-learning algorithm (this development follows that of Ertefaie (2014), Luckett *et al.* (2016) and Meyer *et al.* (2017)). We implicitly assume that the causal conditions and the homogeneous Markov conditions that are stated in the main paper hold; however, we do not assume that the postulated parametric models for the disease dynamics are correctly specified.

For any allocation strategy $d \in \mathcal{D}$ write $\mathbb{E}^d$ to denote expectation with respect to the trajectory distribution that is induced by assigning allocations according to d; an omitted superscript, i.e. $\mathbb{E}$, indicates expectation with respect to the underlying data-generating model. For each time t, allocation strategy $d \in \mathcal{D}$ and state allocation pair $(\mathbf{s}^t, \mathbf{a}^t)$ define the Q-function

$$Q(\mathbf{s}^t, \mathbf{a}^t, d) = \mathbb{E}^d \left\{ \sum_{k \geqslant 0} \gamma^k u(\mathbf{Y}^{t+k}) | \mathbf{S}^t = \mathbf{s}^t \mathbf{A}^t = \mathbf{a}^t \right\},$$

so that $Q(\mathbf{s}^t, \mathbf{a}^t)$ measures the expected discounted utility if the epidemic status is $\mathbf{S}^t = \mathbf{s}^t$, allocation $\mathbf{A}^t = \mathbf{a}^t$ is chosen at time t and allocation strategy d is followed thereafter. Using the law of the iterated expectation (and some algebra), it can be seen that the Q-function satisfies

$$\mathbb{E}([u(\mathbf{Y}^t) + \gamma Q\{\mathbf{S}^{t+1}, d(\mathbf{S}^{t+1}), d\} - Q(\mathbf{S}^t, \mathbf{A}^t, d)]\phi(\mathbf{S}^t, \mathbf{A}^t)) = 0, \tag{8}$$

where ϕ is an arbitrary function defined on dom $\mathbf{S}^t \times \mathbf{A}^t$. The expectation in equation (8) is with respect to the data-generating model and thus will form the basis of an estimating equation. Let $\{Q(\mathbf{s}^t, \mathbf{a}^t, d; \lambda) : \lambda \in \Lambda \subseteq \mathbb{R}^q\}$ be a class of parametric models for $Q(\mathbf{s}^t, \mathbf{a}^t, d)$ which are assumed to be differentiable with respect to λ. Let $\phi(\mathbf{s}^t, \mathbf{a}^t, d) = \nabla_\lambda Q(\mathbf{s}^t, \mathbf{a}^t, d; \lambda)$; then the sample analogue of equation (8) is

$$\sum_{v=1}^{t-1} ([u(\mathbf{Y}^v) + \gamma Q\{\mathbf{S}^{v+t}, d(\mathbf{S}^{v+1}), d; \lambda\} - Q(\mathbf{S}^v, \mathbf{A}^v, d; \lambda)] \nabla_\lambda Q(\mathbf{S}^v, \mathbf{A}^v, d; \lambda)) = 0.$$

Let $\hat{\lambda}^t(d)$ denote a solution to this estimating equation. To obtain a draw from the sampling distribution of $\hat{\lambda}^t(d)$ we use the multiplier bootstrap. As in the preceding section, let $\mathbf{M}^{t-1} = M_1^{t-1}, \ldots, M_{t-1}^{t-1}$ denote IID non-negative random variables with mean 1 and variance 1. Define $\hat{\lambda}^t_{M^t}(d)$ to be the solution to the perturbed estimating equation

$$\sum_{v=1}^{t-1} M_v^{t-1}([u(\mathbf{Y}^v) + \gamma Q\{\mathbf{S}^{v+1}, d(\mathbf{S}^{v+1}), d; \lambda\} - Q(\mathbf{S}^v, \mathbf{A}^v, d; \lambda)]\nabla_\lambda Q(\mathbf{S}^v, \mathbf{A}^v, d; \lambda)) = 0.$$

At each time t, the TS Q-learning estimator draws an allocation strategy

$$\breve{d}^t = \arg\max_{d \in \mathcal{D}} Q\{\mathbf{S}^t, d(\mathbf{S}^t), d; \hat{\lambda}^t_{M^t}\}.$$

As described previously, one could tune the amount of exploration through the variance of the resampling weights.

Because Q-learning imposes less structure on the underlying generative model it may be more robust to misspecification than TS as we proposed it; however, imposing less structure may lead to higher variability especially early in the intervention process where data are scarce. An interesting research direction is the combination of the two methods; for example, one might use the estimated dynamics model to augment the estimating equation or to simulate the exploration–exploitation trade-off.

Extensions to other applications

The methodology proposed was designed to manage the spread of a replicating agent across space and time. Henderson and Lizotte identified novel applications of the methodology in the context of precision medicine. Henderson points out that the exploration–exploitation trade-off arises in adaptive clinical trials and asks whether the methodology might be extended to this setting. Indeed, TS was originally designed for sequential treatment allocation across a stream of IID patients (Thompson, 1933). However, this original formulation of TS considered only a single decision point and did not account for patient covariates. Furthermore, whereas there is a large literature on adaptive clinical trials (see Berry and Fristedt (1985), Thall and Wathen (2007), Berry (2012), Berry *et al*. (2010), Yin (2012) and references therein), the work on sequential adaptive treatment allocation within patients over time is less well developed (see Nahum-Shani *et al*. (2014), Cheung *et al*. (2015) and Klasnja *et al*. (2015)). An extension to IID data based on the Q-learning algorithm that was described above is as follows.

Suppose that we have data on n patients of the form $\{(\mathbf{S}_i^1, \mathbf{A}_i^1, \mathbf{Y}_i^1, \ldots, \mathbf{S}_i^{T_i}, \mathbf{A}_i^{T_i}, \mathbf{Y}_i^{T_i})\}_{i=1}^n$ where $\mathbf{S}_i^t \in \mathbb{R}^p$ is the state of patient i at treatment stage t, $\mathbf{A}_i^t \in \mathcal{A}$ is the treatment that is assigned to patient i at treatment stage t, $\mathbf{Y}_i^t \in \mathbb{R}^q$ is a vector of outcomes for patient i at treatment stage t and T_i is the number of treatment stages that patient i has completed. We assume that the state is Markov and homogeneous in that $P(\mathbf{S}_i^{t+1}|\bar{\mathbf{Y}}_i^i, \bar{\mathbf{A}}_i^t, \bar{\mathbf{S}}_i^t) = P(\mathbf{S}_i^{t+1}|\mathbf{S}_i^t, \mathbf{A}_i^t)$ and that this probability does not depend on the subject i or the time t. Similarly, we assume that $P(\mathbf{Y}_i^t|\bar{\mathbf{Y}}_i^{t-1}, \bar{\mathbf{A}}_i^t, \bar{\mathbf{S}}_i^t) = P(\mathbf{Y}_i^t|\mathbf{S}_i^t, \mathbf{A}_i^t)$ and that this probability does not depend on i or t. Consider a patient presenting with state $\mathbf{S} = \mathbf{s}$; this may be one of the existing $i = 1, \ldots, n$ patients visiting the clinic for the $(T_i + l)$st stage of treatment or a new patient. Let $\mathcal{D}$ denote the set of possible treatment regimes, i.e. deterministic mappings from states to treatments. To assign their treatment by using TS, for each $i = 1, \ldots, n$, draw $\mathbf{M}_i^{T_i-1} = M_{1,i}^{T_i-1}, \ldots, M_{T_i-1,i}^{T_i-1}$ IID non-negative random variables with mean and variance 1 and solve the estimating equation

$$\sum_{i=1}^{n} \sum_{t=1}^{T_i-1} M_{t,i}^{T_i-1}([u(\mathbf{Y}_i^t) + \gamma Q\{\mathbf{S}_i^{t+1}, d(\mathbf{S}_i^{t+1}), d; \lambda\} - Q(\mathbf{S}_i^t, \mathbf{A}_i^t, d; \lambda)]\nabla_\lambda Q(\mathbf{S}_i^t, \mathbf{A}_i^t, d; \lambda)) = 0$$

to obtain $\hat{\lambda}_{n,m}(d)$ and subsequently $\tilde{d}_n = \arg\max_{d \in \mathcal{D}} Q\{\mathbf{S}, d(\mathbf{S}), d; \hat{\lambda}_{n,\mathrm{m}}(d)\}$ and treatment assignment $\tilde{d}(\mathbf{S})$. On observing their response to treatment, the set of available data can be updated and the process repeated at each subsequent treatment decision. Deriving the operating characteristics of this procedure, e.g. power and sample size calculation, is an interesting direction for interesting future work (see Laber *et al*. (2016) for a description of technical issues that are associated with sizing a study for estimation of an optimal treatment regime).

Decision support

Data-driven decision making in health seeks to provide decision support, i.e. to inform decisions rather than to dictate them. Thus, as stated by Lizotte and Thall, the goal should be to build systems that

extract information from accumulating data that is meaningful and actionable in a domain context. The methodology proposed falls short of this goal in that the recommended allocations are the output of a rather complicated algorithm that is attempting to satisfy the dual objective of managing short-term disease spread with information gathering for better long-term control. However, we view the methodology proposed as first steps in building the back-end of a deployable decision support system. As noted by Thall and Lizotte, an effective system would provide an easy-to-use user interface to model-fitting routines and visualizations, as well as qualitative and quantitative descriptions of the fitted models and recommended allocations. We believe that research in this direction is among the most important and most neglected in statistical data-driven decision making.

References in the discussion

Agarwal, A., Bird, S., Cozowicz, M., Hoang, L., Langford, J., Lee, S., Li, J., Melamed, D., Oshri, G., Ribas, O., Sen, S. and Slivkins, A. (2016) Making contextual decisions with low technical debt. *Preprint arXiv:1606.03966*.

Agarwal, A., Hsu, D., Kale, S., Langford, J., Li, L. and Schapire, R. (2014) Taming the monster: a fast and simple algorithm for contextual bandits. In *Proc. Int. Conf. Machine Learning* (eds E. P. Xing and T. Jebara), pp. 1638–1646.

Agrawal, R. (1995) Sample mean based index policies by $O(\log n)$ regent for the multi-armed bandit problem. *Adv. Appl. Probab.*, **27**, 1054–1078.

Aitken, C. K., Lewis, J., Tracy, S. L., Spelman, T., Bowden, D. S., Bharadwaj, M., Drummer, H. and Hellard, M. (2008) High incidence of hepatitis C virus reinfection in a cohort of injecting drug users. *Hepatology*, **48**, 1746–1752.

Audibert, J.-Y., Munos, R. and Szepesvári, C. (2009) Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. *Theoret. Comput. Sci.*, **410**, 1876–1902.

Bastain, H., Bayati, M. and Khosravi, K. (2018) Mostly exploration-free algorithms for contextual bandits. *Preprint arXiv:1704.09011*.

Berry, D. A. (2012) Adaptive clinical trials in oncology. *Nat. Rev. Clin. Oncol.*, **9**, 199.

Berry, S. M., Carlin, B. P., Lee, J. J. and Muller, P. (2010) *Bayesian Adaptive Methods for Clinical Trials*. Boca Raton: CRC Press.

Berry, D. A. and Fristedt, B. (1985) *Bandit Problems: Sequential Allocation of Experiments*. New York: Springer.

Bietti, A., Agarwal, A. and Langford, J. (2018) Practical evaluation and optimization of contextual bandit algorithms. *Preprint arXiv:1802.04064*.

Borth, D. M. (1975) A total entropy criterion for the dual problem of model discrimination and parameter estimation. *J. R. Statist. Soc.* B, **37**, 77–87.

Box, G. E. P. and Hill, W. J. (1967) Discrimination among mechanistic models. *Technometrics*, **9**, 57–71.

Carroll, R. J., Ruppert, D., Crainiceanu, C. M. and Stefanski, L. A. (2006) *Measurement Error in Nonlinear Models: a Modern Perspective*. Boca Raton: Chapman and Hall–CRC.

Chakraborty, B., Ghosh, P., Moodie, E. E. M. and Rush, A. J. (2016) Estimating optimal shared-parameter dynamic regimens with application to a multistage depression clinical trial. *Biometrics*, **72**, 865–876.

Chapelle, O. and Li, L. (2011) An empirical evaluation of Thompson sampling. In *Advances in Neural Information Processing Systems 24* (eds J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira and K. Q. Weinberger), pp. 2249–2257.

Chen, G., Zeng, D. and Kosorok, M. R. (2016) Personalized dose finding using outcome weighted learning. *J. Am. Statist. Ass.*, **111**, 1509–1521.

Chesterton, G. K. (1917) *A Short History of England*. London: Chatto and Windus.

Cheung, Y. K., Chakraborty, B. and Davidson, K. W. (2015) Sequential multiple assignment randomized trial (SMART) with adaptive randomization for quality improvement in depression treatment program. *Biometrics*, **71**, 450–459.

Choi, A. L., Kim, D. W. and Lai, T. L. (2019) *Personalized Recommendation Technology and Health Analytics*. Hoboken: Wiley. To be published.

Coleman, J. S., Katz, E. and Menzel, H. (1957) The diffusion of an innovation among physicians. *Sociometry*, **20**, 253–270.

Cornelison, C. T., Keel, M. K., Gabriel, K. T., Barlament, C. K., Tucker, T. A., Pierce, G. E. and Crow, S. A. (2014) A preliminary report on the contact-independent antagonism of pseudogymnoascus destructans by rhodococcus rhodochrous strain dap96253. *BMC Microbiol.*, **14**, no. 1, article 246.

Cox, D. R. (1958) *Planning of Experiments*. New York: Wiley.

Daraganova, G., Pattison, P., Koskinen, J., Mitchell, B., Bill, A., Watts, M. and Baum, S. (2012) Networks and geography: modelling community network structures as the outcome of both spatial and network processes. *Socl Netwrks*, **34**, 6–17.

Dawid, A. P. and Didelez, V. (2010) Identifying the consequences of dynamic treatment strategies: a decision-theoretic overview. *Statist. Surv.*, **4**, 184–231.

Deardon, R., Brooks, S. P., Grenfell, B. T., Keeling, M. J., Tildesley, M. J., Savill, N. J., Shaw, D. J. and Woolhouse, M. E. (2010) Inference for individual-level models of infectious diseases in large populations. *Statist. Sin.*, **20**, 239.

Dudík, M., Erhan, D., Langford, J. and Li, L. (2014) Doubly robust policy evaluation and optimization. *Statist. Sci.*, **29**, 485–511.

Eckardt, M. and Mateu, J. (2018) Structured regression modeling of network_intensity functions for spatial point patterns. To be published.

Eckles, D. and Kaptein, M. (2014) Thompson sampling with the online bootstrap. *Preprint arXiv:1410.4009*.

Efron, B. (2012) Bayesian inference and the parametric bootstrap. *Ann. Appl. Statist.*, **6**, 1971–1997.

Ertefaie, A. (2014) Constructing dynamic treatment regimes in infinite-horizon settings. *Preprint arXiv:1406.0764*.

Ferreira, M. A. R. (2015) Inhomogeneous evolutionary MCMC for Bayesian optimal sequential environmental monitoring. *Environ. Ecol. Statist.*, **22**, 705–724.

Ferreira, M. A. R. and Sanyal, N. (2014) Bayesian optimal sequential design for nonparametric regression via inhomogeneous evolutionary MCMC. *Statist. Methodol.*, **18**, 131–141.

Geisser, S. (2017) *Predictive Inference*. Abingdon: Routledge.

Ghavamzadeh, M., Mannor, S., Pineau, J. and Tamar, A. (2015) *Bayesian Reinforcement Learning: a Survey*. Singapore: World Scientific Publishers.

Greenan, C. C. (2015) Diffusion of innovations in dynamic networks. *J. R. Statist. Soc.* A, **178**, 147–166.

Guez, A., Silver, D. and Dayan, P. (2014) Better optimism by Bayes: adaptive planning with rich models. *Preprint arXiv:1402.1958*.

Handcock, M. and Gile, K. (2010) Modeling social networks from sampled data. *Ann. Appl. Statist.*, **4**, 5–25.

Hoyt, J. R., Cheng, T. L., Langwig, K. E., Hee, M. M., Frick, W. F. and Kilpatrick, A. M. (2015) Bacteria isolated from bats inhibit the growth of pseudogymnoascus destructans, the causative agent of white-nose syndrome. *PLOS One*, **10**, no. 4, article e0121329.

Jenness, S. M., Goodreau, S. M. and Morris, M. (2016) EpiModel: mathemarical modeling of infectious disease. *R Package Version 1.2.6*. (Available from `http://epimodel.org/`.)

Jewell, C. P., Kypraios, T., Christley, R. and Roberts, G. O. (2009a) A novel approach to real-time risk prediction for emerging infectious diseases: a case study in avian influenza h5n1. *Prev. Veter. Med.*, **91**, 19–28.

Jewell, C. P., Kypraios, T., Neal, P. and Roberts, G. O. (2009b) Bayesian analysis for emerging infectious diseases. *Baysn Anal.*, **4**, 465–496.

Kaelbling, L. P., Littman, M. L. and Cassandra, A. R. (1998) Planning and acting in partially observable stochastic domains. *Artif. Intell.*, **101**, 99–134.

Kaelbling, L. P., Littman, M. L. and Moore, A. W. (1996) Reinforcement learning: a survey. *J. Artif. Intell. Res.*, **4**, 237–285.

Khavarzadeh, R., Mohammadzadeh, M. and Mateu, J. (2018) A simple two-step method for spatio-temporal design-based_balanced sampling. *Stoch. Environ. Res. Risk Assessmnt*, **32**, 457–468.

Klasnja, P., Hekler, E. B., Shiffman, S., Boruvka, A., Almirall, D., Tewari, A. and Murphy, S. A. (2015) Micro-randomized trials: an experimental design for developing just-in-time adaptive interventions. *Hlth Psychol.*, **34**, S, 12–20.

Kosorok, M. R. (2008) *Introduction to Empirical Processes and Semiparametric Inference*. New York: Springer.

Krivitsky, P. N. and Morris, M. (2017) Inference for social network models from egocentrically sampled data, with application to understanding persistent racial disparities in HIV prevalence in the US. *Ann. Appl. Statist.*, **11**, 427–455.

Kypraios, T. and O'Neill, P. D. (2018) Bayesian nonparametrics for stochastic epidemic models. *Statist. Sci.*, **33**, 44–56.

Laber, E. B., Zhao, Y.-Q., Regh, T., Davidian, M., Tsiatis, A., Stanford, J. B., Zeng, D., Song, R. and Kosorok, M. R. (2016) Using pilot data to size a two-arm randomized trial to find a nearly optimal personalized treatment strategy. *Statist. Med.*, **35**, 1245–1256.

Lai, T. L. and Robbins, H. (1985) Asymptotically efficient adaptive allocation rules. *Adv. Appl. Math.*, **6**, 4–22.

Lee, J., Thall, P., Ji, Y. and Müller, P. (2015) Bayesian dose-finding in two treatment cycles based on the joint utility of efficacy and toxicity. *J. Am. Statist. Ass.*, **110**, 711–722.

Little, R. J. and Rubin, D. B. (2014) *Statistical Analysis with Missing Data*. Hoboken: Wiley.

Lu, X. and Van Roy, B. (2017) Ensemble sampling. In *Advances in Neural Information Processing Systems*, pp. 3260–3268.

Luckett, D. J., Laber, E. B., Kahkoska, A. R., Maahs, D. M., Mayer-Davis, E. and Kosorok, M. R. (2016) Estimating dynamic treatment regimes in mobile health using V-learning. *Preprint arXiv:1611.03531*.

Lusher, D., Koskinen, J. and Robins, G. (2013) *Exponential Random Graph Models for Social Networks: Theory, Methods and Applications*. Cambridge: Cambridge University Press.

Maher, S. P., Kramer, A. M., Pulliam, J. T., Zokan, M. A., Bowden, S. E., Barton, H. D., Magori, K. and Drake, J. M. (2012) Spread of white-nose syndrome on a network regulated by geography and climate. *Nat. Communs*, **3**, 1306.

May, B. C., Korda, N., Lee, A. and Leslie, D. S. (2012) Optimistic Bayesian sampling in contextual-bandit problems. *J. Mach. Learn. Res.*, **13**, 2069–2106.

McKinley, T. J., Vernon, I., Andrianakis, I., McCreesh, N., Oakley, J. E., Nsubuga, R. N., Goldstein, M. and White, R. G. (2018) Approximate Bayesian computation and simulation-based inference for complex stochastic epidemic models. *Statist. Sci.*, **33**, 4–18.

Meyer, N., Clifton, J., Laber, E., Pacifici, B., Reich, J. and Drake, J. (2017) *q*-learning for spatio-temporal problems with application to Ebola virus disease in West Africa. *Technical Report*. Department of Statistics, North Carolina State University, Raleigh.

Morris, M. (2004) *Network Epidemiology: a Handbook for Survey Design and Data Collection*. Oxford: Oxford University Press.

Müller, P., Sansó, B. and De Iorio, M. (2004) Optimal Bayesian design by inhomogeneous Markov chain simulation. *J. Am. Statist. Ass.*, **99**, 788–798.

Murphy, S. A. (2005) A generalization error for Q-learning. *J. Mach. Learn. Res.*, **6**, 1073–1097.

Murphy, S. A., van der Laan, M. J., Robins, J. M. and CPPR Group (2001) Marginal mean models for dynamic regimes. *J. Am. Statist. Ass.*, **96**, 1410–1423.

Nahum-Shani, I., Smith, S. N., Tewari, A., Witkiewitz, K., Collins, L. M., Spring, B. and Murphy, S. (2014) Just in time adaptive interventions (JITAIS): an organizing framework for ongoing health behavior support. In *Methodology Center Technical Report, 2014*, pp. 14–126.

Newton, M. A. and Raftery, A. E. (1994) Approximate Bayesian inference with the weighted likelihood bootstrap (with discussion). *J. R. Statist. Soc.* B, **56**, 3–48.

O'Neill, P. D. (2010) Introduction and snapshot review: relating infectious disease transmission models to data. *Statist. Med.*, **29**, 2069–2077.

O'Neill, P. D. and Roberts, G. O. (1999) Bayesian inference for partially observed stochastic epidemics. *J. R. Statist. Soc.* A, **162**, 121–129.

Osband, I., Blundell, C., Pritzel, A. and Van Roy, B. (2016) Deep exploration via bootstrapped DQN. In *Advances in Neural Information Processing Systems 29* (eds D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon and R. Garnett), pp. 4026–4034.

Palmer, J. M., Drees, K. P., Foster, J. T. and Lindner, D. L. (2018) Extreme sensitivity to ultraviolet light in the fungal pathogen causing white-nose syndrome of bats. *Nat. Communs*, **9**, no. 1.

Poupart, P. and Vlassis, N. (2008) Model-based Bayesian reinforcement learning in partially observable domains. In *Proc Int. Symp. Artificial Intelligence and Mathematics*, pp. 1–2.

Praestgaard, J. (1990) Bootstrap with general weights and multiplier central limit theorems. *Technical Report 195*. Department of Statistics, University of Washington, Seattle.

Rich, B., Moodie, E. E. M. and Stephens, D. A. (2016) Optimal individualized dosing strategies: a pharmacologic approach to developing dynamic treatment regimens for continuous-valued treatments. *Biometr. J.*, **58**, 502–517.

Robins, J. M. (2004) Optimal structural nested models for optimal sequential decisions. In *Proc. 2nd Seattle Symp. Biostatistics* (eds D. Lin and P. Heagerty), pp. 189–326. New York: Springer.

Robins, G. L., Pattison, P. E. and Woolcock, J. (2005) Small and other worlds: global network structures from local processes. *Am. J. Sociol.*, **110**, 894–936.

Rolls, D. A., Daraganova, G., Sacks-Davis, R., Hellard, M., Jenkinson, R., McBryde, E., Pattison, P. E. and Robins, G. L. (2012) Modelling hepatitis C transmission over a social network of injecting drug users. *J. Theoret. Biol.*, **297**, 73–87.

Rolls, D. A., Daraganova, G., Sacks-Davis, R., Hellard, M., Jenkinson, R., McBryde, E., Pattison, P. E. and Robins, G. L. (2013a) Modelling a disease-relevant contact network of people who inject drugs. *Socl Netwrks*, **35**, 699–710.

Rolls, D. A., Sacks-Davis, R., Jenkinson, R., McBryde, E., Pattison, P., Robins, G. and Hellard, M. (2013b) Hepatitis C transmission and treatment in contact networks of people who inject drugs. *PLOS One*, **8**, no. 11, article e78286.

Ross, R. (1916) An application of the theory of probabilities to the study of *a priori* pathometry, Part I. *Proc. R. Soc. Lond.* A, **92**, 204–230.

Ross, S., Chaib-draa, B. and Pineau, J. (2008) Bayes-adaptive pomdps. In *Advances in Neural Information Processing Systems*, pp. 1225–1232.

Rubin, D. B. (1984) Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Ann. Statist.*, **12**, 1151–1172.

Ryan, E. G., Drovandi, C. C., McGree, J. M. and Pettitt, A. N. (2016) A review of modern computational algorithms for Bayesian optimal design. *Int. Statist. Rev.*, **84**, 128–154.

Schulte, P., Tsiatis, A., Laber, E. and Davidian, M. (2014) Q- and A-learning methods for estimating optimal dynamic treatment regimes. *Statist. Sci.*, **29**, 640–661.

Shannon, C. E. (1948) A mathematical theory of communication. *Bell Syst. Tech. J.*, **27**, 379–423; 623–656.

Si, J., Barto, A. G., Powell, W. B. and Wunsch, D. C. (2004) *Handbook of Learning and Approximate Dynamic Programming*. Los Alamitos: Institute of Electrical and Electronics Engineers Press.

Simon, G., Johnson, E., Lawrence, J., Rossom, R. C., Ahmedani, B., Lynch, F. M., Beck, A., Waitzfelder, B., Ziebell, R., Penfold, R. B. and Shortreed, S. M. (2018) Predicting suicide attempts and suicide deaths following outpatient visits using electronic health records. *Am. J. Psychiatr.*, to be published.

Stoica, R. S., Philippe, A., Gregori, P. and Mateu, J. (2017) ABC Shadow algorithm: a tool for statistical analysis of spatial patterns. *Statist. Comput.*, **27**, 1225–1238.
Sutton, R. S. and Barto, A. G. (1998) *Reinforcement Learning: an Introduction*. Cambridge: MIT Press.
Sweeting, T. and Kharroubi, S. (2005) Application of a predictive distribution formula to Bayesian computation for incomplete data models. *Statist. Comput.*, **15**, 167–178.
Tao, Y. and Wang, L. (2016) Adaptive contrast weighted learning for multistage multi-treatment decision-making. *Biometrics*, **73**, 145–155.
Tao, Y., Wang, L. and Almirall, D. (2018) Tree-based reinforcement learning for estimating optimal dynamic treatment regimes. *Ann. Appl. Statist.*, to be published.
Thall, P. F. and Wathen, J. K. (2007) Practical Bayesian adaptive randomisation in clinical trials. *Eur. J. Cancer*, **43**, 859–866.
Thompson, W. R. (1933) On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, **25**, 285–294.
Tien, J. H. and Earn, D. J. (2010) Multiple transmission pathways and disease dynamics in a waterborne pathogen model. *Bull. Math. Biol.*, **72**, 1506–1533.
Toni, T., Welch, D., Strelkowa, N., Ipsen, A. and Stumpf, M. P. (2009) Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *J. R. Soc. Interfc.*, **6**, 187–202.
Valente, T. W. (2012) Network interventions. *Science*, **337**, 49–53.
Van der Vaart, A. and Wellner, J. (1996) *Weak Convergence and Empirical Processes: with Applications to Statistics*. New York: Springer.
Verant, M. L., Boyles, J. G., Waldrep, W., Wibbelt, G. and Blehert, D. S. (2012) Temperature-dependent growth of "Geomyces destructans", the fungus that causes bat white-nose syndrome. *PLOS One*, **7**, no. 1, article e46280.
Watkins, C. J. and Dayan, P. (1992) *Q*-learning. *Mach, Learn.*, **8**, 279–292.
West, M. and Harrison, J. (2006) *Bayesian Forecasting and Dynamic Models*. New York: Springer Science and Business Media.
Wiering, M. and van Otterlo, M. (2012) *Reinforcement Learning: State-of-the-art*, vol. 12. New York: Springer.
Willis, C. K. and Brigham, R. M. (2004) Roost switching, roost sharing and social cohesion: forest-dwelling big brown bats, Eptesicus fuscus, conform to the fission–fusion model. *Anim. Behav.*, **68**, 495–505.
Yin, G. (2012) *Clinical Trial Design: Bayesian and Frequentist Adaptive Methods*. Hoboken: Wiley.
Zhang, B., Tsiatis, A. A., Laber, E. B. and Davidian, M. (2012) A robust method for estimating optimal treatment regimes. *Biometrics*, **68**, 1010–1018.
Zhang, B., Tsiatis, A. A., Laber, E. B. and Davidian, M. (2013) Robust estimation of optimal dynamic treatment regimes for sequential treatment decisions. *Biometrika*, **100**, 681–694.
Zhou, X., Wang, Y. and Zeng, D. (2018) Sequential outcome-weighted multicategory learning for estimating optimal individualized treatment rules. To be published.