COMMUN. MATH. SCI.
Vol. 12, No. 5, pp. 859–877

© 2014 International Press

MARKOV CONTROL PROCESSES WITH RARE STATE OBSERVATION: THEORY AND APPLICATION TO TREATMENT SCHEDULING IN HIV-1[*]

STEFANIE WINKELMANN[†], CHRISTOF SCHÜTTE[‡], AND MAX VON KLEIST[§]

Abstract. Markov Decision Processes (MDP) or Partially Observable MDPs (POMDP) are used for modelling situations in which the evolution of a process is partly random and partly controllable. These MDP theories allow for computing the optimal control policy for processes that can continuously or frequently be observed, even if only partially. However, they cannot be applied if state observation is very costly and therefore rare (in time). We present a novel MDP theory for rare, costly observations and derive the corresponding Bellman equation. In the new theory, state information can be derived for a particular cost after certain, rather long time intervals. The resulting information costs enter into the total cost and thus into the optimization criterion. This approach applies to many real world problems, particularly in the medical context, where the medical condition is examined rather rarely because examination costs are high. At the same time, the approach allows for efficient numerical realization. We demonstrate the usefulness of the novel theory by determining, from the national economic perspective, optimal therapeutic policies for the treatment of the human immunodeficiency virus (HIV) in resource-rich and resource-poor settings. Based on the developed theory and models, we discover that available drugs may not be utilized efficiently in resource-poor settings due to exorbitant diagnostic costs.

Key words. information costs, hidden state, Bellman equation, optimal therapeutic policies, diagnostic frequency, resource-poor, resource-rich.

AMS subject classifications. 49N30, 60J27, 60J28, 93B07, 90C40, 93E20.

1. Introduction

Many natural phenomena are intrinsically stochastic and can be accurately described in terms of continuous-time Markov processes [1]. Controlling such processes is of fundamental interest [2]. This is also the case for many medical applications where the medical condition of the patient is governed by a stochastic process, and medical treatment can be interpreted as controlling the medical condition.

The theory of Markov Decision Processes (MDP) is naturally designed to find optimal controls for Markov processes. The standard theory of MDPs, however, relies on several assumptions that restrict their practical use in medical applications. The main restriction results from the fundamental assumption in standard MDP theory that the process is fully observed and that an action can immediately be adapted based on the state observation. In the medical context, however, the state of the system, i.e., the medical condition of the respective patient, *cannot* be observed at all times but only at a very limited number of time points where the (costly) medical examinations happen. That is, observations are *rare* and the observation time points cannot be chosen arbitrarily because the related observation costs are part of the total costs that we want to minimize. Moreover, the respective control action, i.e., the medication, can only be adapted at these rare examination time points.

[*]Received: August 31, 2012; accepted (in revised form): July 13, 2013. Communicated by David Cai.

[†]Dep. of Mathematics and Computer Science, Arnimallee 6, Berlin, D-14195, Germany (stefanie.winkelmann@fu-berlin.de).

[‡]Dep. of Mathematics and Computer Science, Arnimallee 6, Berlin, D-14195, Germany (christof.schuette@fu-berlin.de).

[§]Dep. of Mathematics and Computer Science, Arnimallee 6, Berlin, D-14195, Germany (vkleist@zedat.fu-berlin.de).

Partially Observable MDPs (POMDPs) have been designed for cases in which the process is not fully observed and the uncertainty of a measurement has to be taken into account. The theory of POMDPs, however, does not cover the case of *rare* (and costly) observations. Therefore, we present a novel theory for MDPs with rare observations: State "examinations" are separated by lag times τ of blind progress and decisions can only be taken when the state of the system has been determined (at $t+\tau$). Furthermore, each observation produces costs which enter the cost functional, determining a trade-off between frequent examination and increasing costs.

The new MDP theory can describe many situations. In the medical context, for example, an attending physician makes a patient-specific decision (e.g. medication/drug prescription; next examination time) based on the state of health of the patient that he/she encounters upon examination. Examples include dose adaptation to anti-coagulants, or diabetic treatment and the treatment of many infectious diseases. One particular example is the treatment of human immunodeficiency virus (HIV) infection. HIV-1, if untreated, results in the life-threatening acquired immunodeficiency syndrome (AIDS). Several treatment lines are available, which can suppress the virus and relieve the patient from the symptoms of AIDS. However, there is no cure and the treatment 'drives' the development of drug resistance, which renders it inefficient. While each treatment line is effective against particular patterns of viral strains, therapy must be adapted to counter-act drug-resistance development and to enhance long-term control of the virus. In this work we apply the new MDP theory to find optimal treatment-switching and patient monitoring policies for HIV-1. We aim to find patient-specific examination times and sequences of treatments, thus aiming at optimal *individualized* therapy. Our approach is based on a detailed Markov jump process modelling of the virus infection kinetics that captures the intrinsic stochastic nature of HIV drug resistance development.

Optimal control methods have previously been applied by other research groups in the context of HIV-therapy: Luo et al. [3] and Vargas et al. [4] however treated the underlying system deterministically, which fails to capture the intrinsic stochastic nature of HIV drug resistance development [5] and the time-scales on which drug resistance develops. Furthermore, it does not allow for individualized (patient-specific) treatment optimization. Shechter et al. [6] used MDPs (machine maintenance approaches) to maximize expected residual lifetime under treatment, which allows for patient-specific treatment optimization, but unfortunately does not take into account the "cost of observation" and neither the virological state in a patient.

In the manuscript, we will first introduce the new MDP theory for rare state observation, followed by the derivation of a model of viral adaptation to treatment. We assess optimal therapeutic policies for HIV-1 in resource-rich vs. resource-poor countries by taking actual cost-parameters for the respective settings into account. The derived optimal policies are compared with current standard-of-care treatment. Finally, we evaluate parameter sensitivities with respect to patient survival during cost-optimal policies.

2. Markov control processes with rare state observation

In many applications, particularly in a medical context, the state of a system (i.e. health status) cannot be observed at all times and therefore remains hidden to the controller, which contradicts the fundamental assumption of conventional MDPs. Moreover, observation (examination) itself may produce costs and likewise, the action (medication) can only be adapted at a very limited number of time points.

In our example we assume that the action can only be changed when the state has been examined. However, each observation of the systems produces "information costs" k_{info}. As a consequence, finding the optimal control policy includes finding optimal examination times. We will specify an appropriate cost criterion, comprising *process costs* and *information costs*, and formulate the corresponding Bellman equation. Motivated by our medical application (HIV treatment scheduling), we will in the following derive the novel theory for "Markov Control Processes with rare State Observation"and apply it in the subsequent sections.

2.1. The control model. As in a conventional Markov Control Model we consider a set $\mathcal{S}$ of states which, in our application, describe the health status of a patient, as well as a set $\mathcal{A}$ of actions referring to the treatments (drug combinations) that are available in order to control the disease process. We assume both $\mathcal{S}$ and $\mathcal{A}$ to be finite sets. For each action $a \in \mathcal{A}$ there is a generator L_a specifying the dynamics of the process based on action a. More precisely, $L_a(x, y) \geq 0$ is the transition rate for a transition from $x \in \mathcal{S}$ to $y \in \mathcal{S}$, $y \neq x$, given action a, and it holds that $L_a(x, x) = -\sum_{y \neq x} L_a(x, y)$. Furthermore, there is a cost function $c : \mathcal{S} \times \mathcal{A} \to [0, \infty)$ denoting the costs produced by the process per unit of time depending on the actual state and the chosen action. In the medical context, the cost function comprises both the direct costs of the treatment and the indirect costs produced by the health damage of a patient.

Now we assume that the corresponding Markov Control Process $(X_t)_{t \geq 0}$, which is itself continuous in time, may only be observed at discrete (but variable) time points and that each observation produces costs. In this regard, we introduce the novel parameter

$$k_{\text{info}} \geq 0$$

which we will call *cost of information.* This constant fee has to be paid each time that the state of the process is determined, e.g. by a medical examination. As explained in Section 2.2, the total observation costs enter the considered cost criterion such that the frequency of observations cannot be arbitrary. Consequently, finding an optimal control policy means not only finding the optimal action (as in the conventional Markov Control Theory) but also the optimal *examination lag time* $\tau(x) > 0$ for each state (i.e. the amount of time before the next examination takes place).

Since an optimal policy is generally stationary and deterministic in conventional Markov Control Theory [7], we restrict ourselves to stationary and deterministic policies herein.

Controlling the process proceeds according to the following structure: Starting with some known state $X_0 \in \mathcal{S}$ at some *examination time* $t_0 \geq 0$ one chooses an action $a \in \mathcal{A}$ as well as an examination lag time $\tau(X_0) > 0$ defining the next examination time $t_1 = t_0 + \tau(X_0) > t_0$. During the time interval $(t_0, t_1]$ the (random) behavior of the process (X_t) is fully described by the infinitesimal generator L_a and produces costs according to the cost function $c(\cdot, a)$. We do not observe this behavior but only determine the state X_{t_1} of the process at time t_1. For this information expenses k_{info} accrue. Knowing the new state X_{t_1} at time t_1, we choose again an action and a lag time and the procedure restarts.

This is exactly the structure of a medical therapy: By examining the health status of a patient, an appropriate medication is chosen and an appointment for the next examination is arranged.

The described procedure motivates the following definition of the term policy. In the definition, we allow the lag time τ to be infinite, which means that there will be no further examinations.

Definition 2.1 (Policy). *A policy is a function*

$$u : \mathcal{S} \to \mathcal{A} \times (0, \infty], \quad x \mapsto u(x) = \big(a(x), \tau(x)\big)$$

giving for each state $x \in \mathcal{S}$ both an action $a \in \mathcal{A}$ and an examination lag time $\tau > 0$. The set of all such policies is denoted by $\mathcal{U}$.

Remark 2.2. In our setting, the action is fixed for the whole time interval $[t_0, t_0 + \tau)$ and cannot be changed blindly, i.e. without determining the state. This way, the controlled process can be turned into a completely observable discrete Markov Decision Process with random time steps.

Although the policy is deterministic, the actual examination times $(t_k)_{k \in \mathbb{N}}$ are random variables: They are determined by the succession of lag times τ, which by themselves depend on the random states of the process $(X_{t_k})_{k \in \mathbb{N}_0}$. I.e., the points in time where the process is observed are not fixed in advance but vary with the stochastic evolution of the control process.

The procedure is schematically illustrated in figure 2.1: Initially, an optimal (stationary, deterministic) control policy has been derived, which assigns an action $a(x)$ and an examination lag time $\tau(x)$ to each state x. The derived optimal policy is then applied to a stochastic realization of the disease process, e.g., the virus kinetics of an individual patient. Thus, an individualized therapy is implemented in which the examination times t_j are patient-specific while the policy is global, i.e., identical for all patients. The dashed blue lines in figure 2.1 depict the individual, stochastic disease process. The solid dots denote the examination times t_j. As can be seen, at each examination time t_j, a treatment $a(x_{t_j})$ (arrows depicted above the figure) and an examination time lag $\tau(x_{t_j})$ (depicted on the x-axis) is assigned according to the optimal (stationary, deterministic) control law. The treatment is maintained at least for the period $t_j + \tau(x_{t_j})$, irrespective of the (hidden) dynamics in between.

2.2. Cost criterion. The considered time horizon of the cost criteria depends on the particular application at hand. We decided to consider discounted costs on an infinite time horizon. In this setting, the costs arising at time $t > 0$ are weighted by a discount factor $0 < e^{-\lambda t} < 1$, where $\lambda > 0$ is a given constant. The discount factor thus guarantees convergence of the cost functional. For the intended medical treatment application this criterion is suitable because the therapy of a chronic disease does not have a previously known endpoint (i.e. time of death). At the same time, a differentiated weighting of immediate and later costs is reasonable due to an upper limitation of life expectancy and aspects of inflation. In this regard, the concrete choice of the constant λ will depend on the presumed annual inflation in the considered countries; see table 3.2.

Definition 2.3 (Expected Discounted Cost Criterion). *Given a policy $u \in \mathcal{U}$, an initial state $x \in \mathcal{S}$ and a discount factor $\lambda > 0$, the expected discounted cost is defined by*

$$J(x, u) := \mathbb{E}_x^u \left(\sum_{j=0}^{\infty} \left(\int_{t_j}^{t_{j+1}} e^{-\lambda s} c\big(X_s, a(X_{t_j})\big) ds + e^{-\lambda t_{j+1}} k_{\text{info}} \right) \right),$$

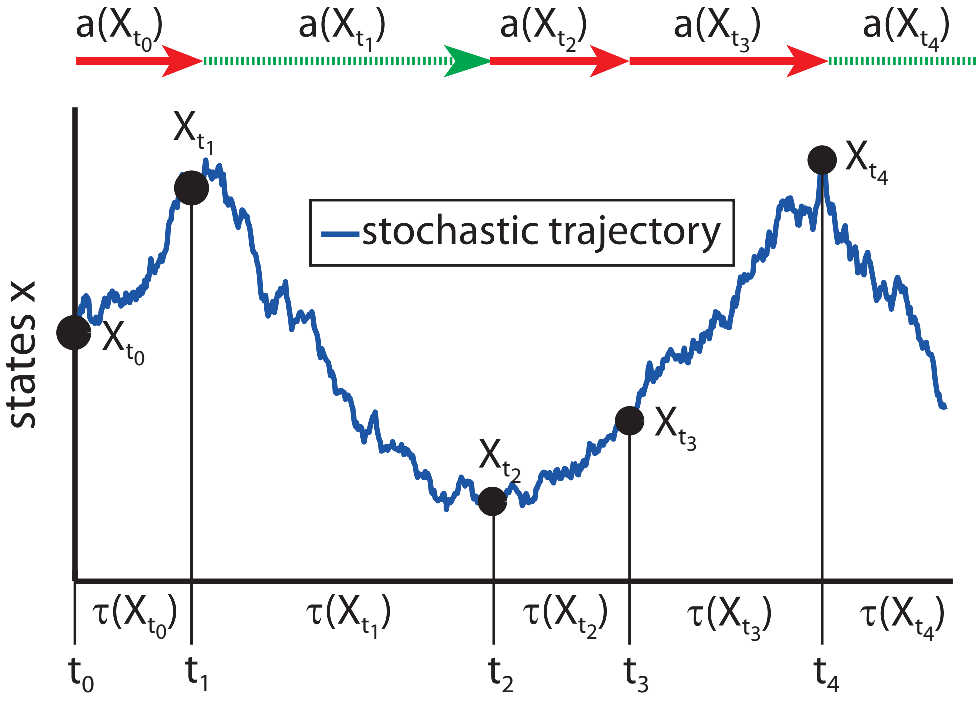

FIG. 2.1. *Schematic illustration of the presented control model. The graphic highlights that the examination times t_j and sequence of actions are random variables because the underlying process is a random realization: The solid blue line indicates the disease process within an individual patient (a stochastic realization of the process), while the solid black dots indicate the examination times t_j for this particular stochastic realization, which depend on the observed state from the last examination time and the stationary deterministic decision rule.*

where $\mathbb{E}_x^u$ stands for the expectation value with respect to the measure determined by x and u. The corresponding value function is

$$V(x) := \inf_{u \in \mathcal{U}} J(x, u).$$

REMARK 2.4. Note that in the cost functional $J(x, u)$, the function c is evaluated in the first argument at X_s with s running in time, while in the second argument it is evaluated at $a(X_{t_j})$ with t_j fixed for each interval. This follows from the fact that the state, which we do not observe during such an interval, may change, while the action stays the same.

In order to simplify the notation and to discover the structure of the given control process we define a new cost function $C : \mathcal{S} \times \mathcal{A} \times [0, \infty] \to [0, \infty)$ by

$$C(x, a, \tau) := \mathbb{E}_x^a \left(\int_0^\tau e^{-\lambda s} c(X_s, a) ds \right). \tag{2.1}$$

These are the expected discounted costs for the time interval $(0, \tau]$ when starting in state $x \in \mathcal{S}$ and choosing action $a \in \mathcal{A}$.

Now, the cost criterion can be written as

$$J(x,u) = \mathbb{E}_x^u \left(\sum_{j=0}^{\infty} e^{-\lambda t_j} \Big(C\big(X_{t_j}, a(X_{t_j}), \tau(X_{t_j})\big) + e^{-\lambda \tau(X_{t_j})} k_{\text{info}} \Big) \right).$$

This notation shows the discrete nature of the system: The process is evaluated only at the evaluation times $t_0, t_1, t_2, ...$ However, as mentioned above, these points are random variables. They depend on the behavior of the process and on the chosen policy. Hence, the process is not equivalent to a usual discrete time control process.

Furthermore, we introduce the operator

$$T_{a,\tau} : \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}^{|\mathcal{S}|}, \quad T_{a,\tau} v := e^{L_a \tau} \cdot v,$$

which is the transition matrix on $\mathcal{S}$ for some fixed time τ and action a. This operator is relevant for the Bellman equation which will be derived in the following section. Together with the cost function C defined in (2.1), the operator $T_{a\tau}$ will permit a clear and compact notation.

2.3. The Bellman equation. Given the cost criterion J we will now formulate a Bellman equation for the corresponding value function $V(\cdot) = \inf_{u \in \mathcal{U}} J(\cdot, u)$. This is the basic step for all numerical computations which follow. We will proceed as follows: Instead of starting with the given optimal control problem and deducing the correct Bellman equation, we formulate a (reasonable) Bellman equation and show that its solution coincides with the value function of our optimal control problem and that the control policy obtained from this Bellman equation is indeed optimal. This approach is common practice (see e.g. [8, 7]), and called *verification*.

THEOREM 2.5 (Verification: Bellman Equation for Discounted Costs). *Assume that $V : S \to \mathbb{R}$ satisfies the Bellman equation*[1]

$$V(x) = \min_{a \in \mathcal{A}, \tau \in [0,\infty]} \Big(C(x,a,\tau) + e^{-\lambda\tau} \Big(k_{\text{info}} + T_{a,\tau} V(x) \Big) \Big), \tag{2.2}$$

and define $\Big(a^*(x), \tau^*(x)\Big) = \arg\min_{a,\tau} \Big(C(x,a,\tau) + e^{-\lambda\tau} \Big(k_{\text{info}} + T_{a,\tau} V(x) \Big) \Big)$. *Then the policy u^* given by $u^*(x) = \big(a^*(x), \tau^*(x)\big)$ fulfills $J(x,u^*) \leq J(x,u)$ for all policies u, and it holds that*

$$J(x,u^*) = V(x). \tag{2.3}$$

The proof can be found in Appendix A.

In many situations (as in our application), the cost function $c : \mathcal{S} \times \mathcal{A} \to [0,\infty)$ is of the form

$$c(x,a) = c_{\mathcal{S}}(x) + c_{\mathcal{A}}(a),$$

with $c_{\mathcal{S}} : \mathcal{S} \to [0,\infty)$ giving the costs produced by the state, e.g. indirect costs produced by the health status of a patient, and $c_{\mathcal{A}} : \mathcal{A} \to [0,\infty)$ denoting the costs produced by the action, e.g. treatment costs. This means that state and action costs

[1] For $\tau = \infty$ the right hand side of equation (2.2) is given by $C(x,a,\infty)$ (see equation (2.1)), which is consistent with the fact that $e^{-\lambda\tau} \overset{\tau\to\infty}{\longrightarrow} 0$ while $k_{\text{info}} + T_{a,\tau}V(x)$ is bounded.

are independent of each other which, especially in our application, is a reasonable assumption. Now, by the linearity of the expectation value, the valuefunction $V(x) = J(x, u^*)$ may also be decomposed by writing

$$V(x) = V_{\mathcal{A}}(x) + V_{\mathcal{S}}(x) + V_{\text{info}}(x),$$

with

$$V_{\mathcal{A}}(x) = \mathbb{E}_x^{u^*}\left(\sum_{j=0}^{\infty}\int_{t_j}^{t_{j+1}} e^{-\lambda s} c_{\mathcal{A}}\big(a(X_{t_j})\big)ds\right),$$

$$V_{\mathcal{S}}(x) = \mathbb{E}_x^{u^*}\left(\sum_{j=0}^{\infty}\int_{t_j}^{t_{j+1}} e^{-\lambda s} c_{\mathcal{S}}\big(X_s\big)ds\right),$$

and

$$V_{\text{info}}(x) = \mathbb{E}_x^{u^*}\left(\sum_{j=0}^{\infty} e^{-\lambda t_{j+1}} k_{\text{info}}\right).$$

For the purpose of interpretation it will be interesting to compute such a cost-splitting, see table 4.2 and its interpretation in Section 4.

Remark 2.6. For the numerical computation of an optimal policy we will use an adapted standard policy iteration algorithm [9]. Compared to the case of full information our setting leads to an extension of the action space while the state space is untouched. The “new” action space is given by

$$\mathcal{A} \times [0, \infty],$$

where the part $[0, \infty]$ needs to be discretized and limited in order to get a numerically feasible set. As for real world applications this is a justifiable step: Depending on the specific application there may be a lower- and upper examination lag time $\tau_{\min}$ and $\tau_{\max}$, respectively. In our application, we set $\tau_{\min} = 1$ days and $\tau_{\max} = 5000$ days in order to numerically compute the optimal policy via the policy iteration algorithm. As the runtime of the policy iteration algorithm grows strongly with an increasing number of states [10], we will formulate a lumped HIV-Model, with a relatively small number of states, that is able to capture the essential features of antiviral therapy and drug resistance development.

3. HIV dynamics model

In the following, we introduce the HIV-model, for which we will apply the previously developed theory.

In Section 3.1 the state space $\mathcal{S}$ and the action space $\mathcal{A}$ of the HIV-model are introduced and in Section 3.2 it is explained how distinct treatments $a \in \mathcal{A}$ manipulate the entries of the generators L_a. In Section 3.3, we will parametrize the corresponding cost functions c for the HIV-model.

3.1. State & action space.

HIV dynamics and -drug resistance development can accurately be described by stochastic reaction kinetics [5, 11, 12]. The fundamental evolution equation for stochastic kinetics is the chemical master equation (CME), for which each state comprises a combination of discrete numbers of individuals of the respective species (e.g. viral strains), resulting in state space dimensions $\mathbb{N}_0 \times \mathbb{N}_0 \times ... \times \mathbb{N}_0$, which is numerically infeasible in terms of a direct solution.

In order to reduce the state-space dimensionality, we introduce a model with a limited number of so-called lumped states, that is motivated by a mechanistic HIV dynamic model [11]: Our model contains four of these lumped states for each virus type: If the respective virus type is absent we denote the respective state by 0, if it is present in low copy numbers, i.e., for < 50 virus copies/mL blood (detection limit of assays used in the clinic), the respective state is denoted by ℓ, for medium copy numbers between 50 and 4000 virus copies/mL blood we denote the lumped states by m and for high copy numbers with more than 4000 virus copies/mL blood, it is h. This coarse graining is in line with the levels of virus produced in the distinct cellular reservoirs of HIV; see e.g. [12]. The ℓ-states are reflecting states, which are justified by inability to eradicate HIV (the persistence of virus in reservoirs [13, 14]), and the h-states are reflecting states, because there is a maximum carrying capacity of the system (i.e. virus does not grow indefinitely). Further, the ℓ-states do not affect patient health (thus not producing costs) as the virus is essentially suppressed [15]. Costs are produced by the h-states and the m-states, respectively, but the h-states produce more costs than the m-states (denoted later in table 3.2).

According to their treatment susceptibility, our model distinguishes 4 viral strains M ("mutants"): a strain WT (wild type) that is susceptible to all treatment lines, a strain R1 which is susceptible to treatment 2 (a_2), but unaffected by (resistant to) treatment 1 (a_1), a strain R2 that is susceptible to a_1, but unaffected by a_2 and a highly resistant strain HR which is resistant to all treatments (a_1&a_2).

Considering all permutations of viral strains $M \in \{\mathrm{WT}, \mathrm{R1}, \mathrm{R2}, \mathrm{HR}\}$ and respective copy numbers $n_C(M) \in \{0, \ell, m, h\}$ and patient death $\maltese$, the state space of the corresponding Markov Control Model turns out to be $\mathcal{S} = \{0, \ell, m, h\}^4 \cup \maltese$ with $|\mathcal{S}| = 4^4 + 1 = 257$ states in total.

In order to describe a state $x \in \mathcal{S}$ we choose a compact vector notation of the form

$$x = \big[n_C(\mathrm{WT})\,,\, n_C(\mathrm{R1}),\, n_C(\mathrm{R2}),\, n_C(\mathrm{HR})\big].$$

For example, the state $x = \big[h\,,\,\ell\,,\,m\,,\,0\big]$ describes the situation of a $h$igh number of wild type strains, a ℓow number of R1-mutants, a $m$edium number of R2-mutants and the absence of highly resistant mutants. We use this notation as well for sets of states by writing, e.g., $\big[\{m, h\}\,,\,*\,,\,0\,,\,0\big]$, which stands for a $m$edium **or** $h$igh number of wild type strains, an arbitrary number of R1-mutants and the absence of R1-mutants and highly resistant mutants.

For the action space we choose the set of treatments $\mathcal{A} = \{a_\emptyset, a_1, a_2\}$, where $a_\emptyset$ denotes the absence of medical intervention, while a_1 and a_2 denote the application of two distinct treatment lines.

3.2. Generator entries. The basic transitions between copy number states $n_C(M)$ for a particular viral strain M (here exemplified for the the wild type strain WT) of our continuous-time Markov model are shown below.

$$\big[\ell\,,*\,,*\,,*\big] \xrightarrow{k_{\ell,a}} \big[m\,,*\,,*\,,*\big], \qquad \big[m\,,*\,,*\,,*\big] \xrightarrow{k_{m,a}} \big[h\,,*\,,*\,,*\big], \tag{3.1}$$

$$\big[m\,,*\,,*\,,*\big] \xrightarrow{\delta_m} \big[\ell\,,*\,,*\,,*\big], \qquad \big[h\,,*\,,*\,,*\big] \xrightarrow{\delta_h} \big[m\,,*\,,*\,,*\big], \tag{3.2}$$

$$\big[h\,,*\,,*\,,*\big] \xrightarrow{d_h} \maltese, \qquad \big[m\,,*\,,*\,,*\big] \xrightarrow{d_m} \maltese, \tag{3.3}$$

$$\big[\ell\,,*\,,*\,,*\big] \xrightarrow{d_\ell} \maltese. \tag{3.4}$$

The parameters $k_{\ell,a}$ and $k_{m,a}$ denote the reaction propensities of going from copy number ℓ to copy number m and from copy number m to copy number h, respectively (viral growth), which depend on the treatment $a \in \{a_\emptyset, a_1, a_2\}$. The parameters δ_m and δ_h are independent of the treatment and denote the reaction propensities for going from copy number m to copy number ℓ and from copy number h to copy number m, respectively (virus elimination). The parameters $d_h > d_m > d_\ell$ denote the propensity for the death of the patient. These parameters are unaffected by the treatments, as well. We assume that high viral burden (states h and m respectively) increases the risk of death, whereas d_ℓ equals the propensity for "natural death". The propensity for natural death was computed according to $d_\ell = 1/(\text{residual life expectancy healthy})$, and is exemplified in the caption of table 3.2. Analogously, d_h and d_m were computed using the average residual life expectancy in states h and m. The following transitions between viral strains M were considered:

$$\big[h, 0, *, *\big] \overset{\mu_{h,\mathrm{R1},a}}{\longrightarrow} \big[h, \ell, *, *\big], \quad \big[m, 0, *, *\big] \overset{\mu_{m,\mathrm{R1},a}}{\longrightarrow} \big[m, \ell, *, *\big], \tag{3.5}$$

$$\big[h, *, 0, *\big] \overset{\mu_{h,\mathrm{R2},a}}{\longrightarrow} \big[h, *, \ell, *\big], \quad \big[m, *, 0, *\big] \overset{\mu_{m,\mathrm{R2},a}}{\longrightarrow} \big[m, *, \ell, *\big], \tag{3.6}$$

$$\big[0, h, *, *\big] \overset{\mu_{h,\mathrm{R1},a}}{\longrightarrow} \big[\ell, h, *, *\big], \quad \big[0, m, *, *\big] \overset{\mu_{m,\mathrm{R1},a}}{\longrightarrow} \big[\ell, m, *, *\big], \tag{3.7}$$

$$\big[0, *, h, *\big] \overset{\mu_{h,\mathrm{R2},a}}{\longrightarrow} \big[\ell, *, h, *\big], \quad \big[0, *, m, *\big] \overset{\mu_{m,\mathrm{R2},a}}{\longrightarrow} \big[\ell, *, m, *\big], \tag{3.8}$$

$$\big[*, h, *, 0\big] \overset{\mu_{h,\mathrm{R2},a}}{\longrightarrow} \big[*, h, *, \ell\big], \quad \big[*, m, *, 0\big] \overset{\mu_{m,\mathrm{R2},a}}{\longrightarrow} \big[*, m, *, \ell\big], \tag{3.9}$$

$$\big[*, *, h, 0\big] \overset{\mu_{h,\mathrm{R1},a}}{\longrightarrow} \big[*, *, h, \ell\big], \quad \big[*, *, m, 0\big] \overset{\mu_{m,\mathrm{R1},a}}{\longrightarrow} \big[*, *, m, \ell\big], \tag{3.10}$$

$$\big[*, *, 0, h\big] \overset{\mu_{h,\mathrm{R1},a}}{\longrightarrow} \big[*, *, \ell, h\big], \quad \big[*, *, 0, m\big] \overset{\mu_{m,\mathrm{R1},a}}{\longrightarrow} \big[*, *, \ell, m\big], \tag{3.11}$$

$$\big[*, 0, *, h\big] \overset{\mu_{h,\mathrm{R2},a}}{\longrightarrow} \big[*, \ell, *, h\big], \quad \big[*, 0, *, m\big] \overset{\mu_{m,\mathrm{R2},a}}{\longrightarrow} \big[*, \ell, *, m\big]. \tag{3.12}$$

The parameters $\mu_{h,\mathrm{R1},a}$ and $\mu_{h,\mathrm{R2},a}$ denote the propensity for the emergence and disappearance of a mutation that confers drug resistance to treatment 1 or 2 (a_1, a_2), respectively, emanating from copy number state h. Analogously, $\mu_{m,\mathrm{R1},a}$ and $\mu_{m,\mathrm{R2},a}$ denote the propensity for the emergence and disappearance of a mutation emanating from copy number states m. Note, that we consider only the following mutations: WT ↔ R1, WT ↔ R2, R1 ↔ HR, and R2 ↔ HR, which is motivated by the fact that a direct transition from WT ↔ HR is very unlikely, because the genetic distance between the two viral strains is too large to be overcome at once. The model is graphically illustrated in figure 3.1.

param.	value	param.	value	param.	value
δ_h	$6.13 \cdot 10^{-2}$	$\mu_{h,\mathrm{R1},\emptyset}$	1.24	$\eta(a_1, \{\mathrm{WT}, \mathrm{R2}\})$	0.979
δ_m	$5.1 \cdot 10^{-2}$	$\mu_{m,\mathrm{R1},\emptyset}$	$4.34 \cdot 10^{-2}$	$\eta(a_1, \{\mathrm{R1}, \mathrm{HR}\})$	0
$k_{\ell,\emptyset}$	0.13	$\mu_{h,\mathrm{R2},\emptyset}$	$2.41 \cdot 10^{-4}$	$\eta(a_2, \{\mathrm{WT}, \mathrm{R1}\})$	0.966
$k_{m,\emptyset}$	0.13	$\mu_{m,\mathrm{R2},\emptyset}$	$2.33 \cdot 10^{-2}$	$\eta(a_2, \{\mathrm{R2}, \mathrm{HR}\})$	0

TABLE 3.1. ***General model parameters.*** *All parameters in units [1/day] except η [unit less].*

The effect of the treatments a_1 and a_2 on the growth and mutation rates is considered in the following way:

$$k_{\ell,a} = \Big(1 - \eta(a, M)\Big) k_{\ell,\emptyset}, \qquad k_{m,a} = \Big(1 - \eta(a, M)\Big) k_{m,\emptyset}, \tag{3.13}$$

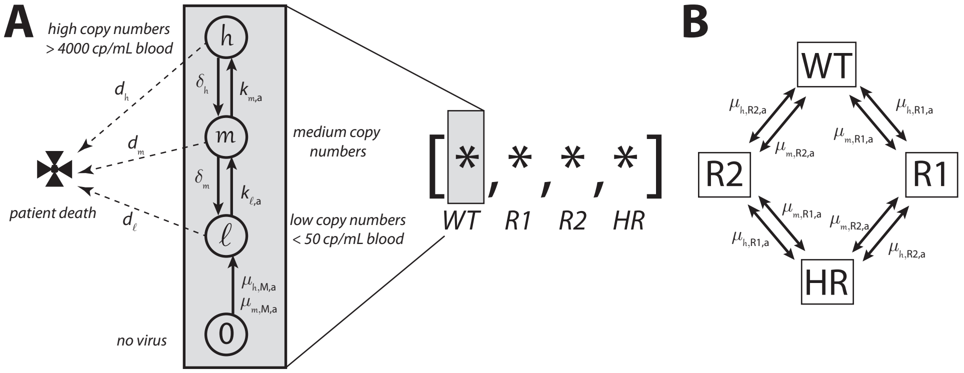

FIG. 3.1. **Simplified HIV Model** *A: Transitions between copy number states n_C. B: Transitions in between viral strains M.*

	Germany	S. Africa		
param.	value	value	unit	ref.
$c_{\mathcal{A}}(a_\emptyset)$	0	0	-	-
$c_{\mathcal{A}}(a_1)$	48.5	0.3	US$/d	[18, 19]
$c_{\mathcal{A}}(a_2)$	58.8	1.08	US$/d	[18, 19]
k_{info}	400	500	US$/d	[16]
d_ℓ	$6.2 \cdot 10^{-5}$	$9.4 \cdot 10^{-5}$	1/d	♮
d_m	$2.7 \cdot 10^{-4}$	$2.7 \cdot 10^{-4}$	1/d	†
d_h	$5.5 \cdot 10^{-4}$	$5.5 \cdot 10^{-4}$	1/d	†
GDP	43,742	8,066	US$/p.p./y	[20]
pL(ℓ)	0	0	-	[16]
pL(m)	0.1	0.1	-	[16]
pL(h)	0.4	0.4	-	[16]
pL(✠)	1	1	-	-
λ	$1 \cdot 10^{-4}$	$1.75 \cdot 10^{-4}$	1/d	‡

TABLE 3.2. ***Country specific model parameters.*** *k_{info} refers to the price for a drug resistance test. The GDP refers to the estimation for the recent year 2011 by the International Monetary Fund. ♮ Computed from the overall residual life expectancy (overall residual life expectancy = overall life expectancy - age of HIV detection), normalized by AIDS related death $\Big($residual life expectancy healthy = (overall residual life expectancy - prevalence*residual life expectancy AIDS)/(1-prevalence)$\Big)$; with d_ℓ = (residual life expectancy healthy)$^{-1}$. The overall life expectancy in Germany and South Africa is 79.4 and 49.3 years, respectively, with an average age of HIV detection of 35 and 25 years and respective HIV prevalence of 0.1% and 17.8%, respectively. † For states m and h we assumed a respective residual life expectancy of 10 and 5 years. ‡ Assuming an annual inflation of 3.5% and 6.2% for Germany and South Africa, respectively.*

$$\mu_{h,\tilde{M},a} = \Big(1 - \eta(a, M)\Big)\mu_{h,\tilde{M},\emptyset}, \quad \mu_{m,\tilde{M},a} = \Big(1 - \eta(a, M)\Big)\mu_{m,\tilde{M},\emptyset}, \qquad (3.14)$$

where $M \in \{\text{WT}, \text{R1}, \text{R2}, \text{HR}\}$ denotes the strain of the reactant virus.The parameter $\eta(a, M)$ denotes the efficacy of treatment a on this viral strain M; i.e. if strain M is susceptible to treatment $a \in \{a_1, a_2\}$, then $0 < \eta(a, M) \leq 1$, and if the viral strain M is insusceptible to treatment $a \in \{a_1, a_2\}$, then $\eta(a, M) = 0$. In the absence of medical intervention $a = a_\emptyset$, $\eta(a, M) = 0$. Therefore, the parameters $k_{\ell,\emptyset}$, $k_{m,\emptyset}$, $\mu_{h,\tilde{M},\emptyset}$, and $\mu_{m,\tilde{M},\emptyset}$ denote the growth rates and respective mutation rates in copy number states m and h in the absence of intervention, i.e. for $a = a_\emptyset$ (see table 3.1).

We estimated all parameters by fitting the presented model to clinical data of virus decay- and rebound dynamics as exemplified in Appendix B.

3.3. Cost parameters. Our analysis is conducted from a country's public health-care/monetary perspective. The costs $c_\mathcal{S}(x)$ of being in the respective states $x \in \mathcal{S}$ were computed based on the average productivity loss $\mathrm{pL}(n_C)$ times the average daily monetary contribution of one individual (assessed in terms of daily per capita GDP), i.e. $c_\mathcal{S}(x) = \mathrm{pL}(x) \cdot \mathrm{GDP}$, with $\mathrm{pL}(x) = \max\limits_{n_C} \mathrm{pL}(n_C)$, where death is interpreted in terms of a complete loss in productivity. All other productivity measures were adapted from [16]. We want to assess optimal policies in the case where two treatment lines (a_1 and a_2 respectively) were available in (i) developed countries, with Germany as a representative, and (ii) in resource-constrained settings, exemplified for South Africa, because of the extraordinary high prevalence (17.8%) of HIV in this country [17]. The daily costs $c_\mathcal{A}$ of treatment in Germany and in resource-constrained settings were derived from [18, 19]. In resource-constrained settings, the William J. Clinton Foundation has negotiated prices for antivirals, which are highly subsidized, allowing access to antivirals in these settings. The respective parameters are displayed in table 3.2. The costs k_info for drug resistance testing are $\approx$ 400 US$ per test in the western world [16]. In resource-constrained settings, these tests are not subsidized. Furthermore, because of the often undeveloped infrastructure tests may be even more expensive (we used the value of $k_\text{info} = 500$ US$ per test). All parameters related to costs are displayed in table 3.2.

4. Application to treatment scheduling in HIV-1

In the following, we will apply the developed Markov control theory of Section 2 to treatment scheduling and diagnostic testing in HIV-1 using the model presented in Section 3. We will first analyze the computed optimal control policy and compare it with current standard-of-care treatment in Section 4.1, before evaluating parameter sensitivities for the optimal control policy with regard to improved patient survival in Section 4.2.

4.1. Action rules & cost allocation. Table 4.1 displays the resulting optimal control policy for the resource-rich and the resource-poor setting given the cost parameters of table 3.2. The composition of the valuefunction and an analysis of its dependence on the cost parameters is computed subsequently.

It can be seen from table 4.1 that treatment 1 (a_1) is chosen whenever (i) only wild type (WT) virus is present, or (ii) when wild type (WT) and strains resistant to treatment 2 (R2) coincide. Treatment 2 (a_2) is only chosen when drug resistance to treatment 1 has emerged (R1), while the virus is still susceptible to treatment 2. Interestingly, there is a difference in the handling of the other states (i.e. highly resistant strains HR, or the concurrence of R1 and R2): While in the context of Germany, no treatment ($a_\emptyset$) is given, treatment 1 (a_1) is applied in the resource-poor setting. This result is due to the fact that the use of treatment in patients that carry drug resistant viruses may provide limited benefit in comparison to the treatment costs for Germany, whereas costs for treatment in resource-constrained settings are in fact so low that their application in the case of drug-resistant virus is still cost-optimal. This assumption is also supported by the cost-splitting in table 4.2 (baseline parameters in first row): For Germany, treatment cost $c_\mathcal{A}$ produce $> 20\%$ of the total

states	action	τ	action	τ
	Germany		**South Africa**	
$\big[\ell\,,0\,,0\,,0\big]$	a_1	155	a_1	$\geq \tau_{\max}$
$\big[\{m,h\}\,,0\,,0\,,0\big]$	a_1	$6-24$	a_1	$11-45$
$\big[*\,,0\,,\{\ell,m,h\}\,,0\big]$	a_1	$20-554$	a_1	$\geq \tau_{\max}$
$\big[*\,,\{\ell,m,h\}\,,0\,,0\big]$	a_2	$159-567$	a_2	$\geq \tau_{\max}$
otherwise	$a_\emptyset$	$\geq \tau_{\max}$	a_1	$\geq \tau_{\max}$

TABLE 4.1. ***Optimal policy.*** *Calculated optimal policy for the resource-rich (Germany) and resource-poor settings (South Africa) giving the treatment and the examination lag time τ (in days) depending on the state of the patient. For clarity reasons, states are merged according to their related treatment choice. The values given for τ refer to the respective minimum and maximum value of $\tau(x)$ for the states x indicated in the first column; e.g. for the second row it holds that $\tau(\big[h\,,0\,,0\,,0\big])=6$ and $\tau(\big[m\,,0\,,0\,,0\big])=24$ for Germany.*

costs, whereas they only produce about 2.5% of the total costs in South Africa. In fact, Stoll et al. [18] argued that treatment may be too expensive in Germany, because of the use of original manufacturer's drugs instead of generic drugs.

As can be seen in table 4.1, much longer periods between tests are proposed in the resource-constrained setting in comparison to the resource-rich setting. In fact, drug-resistance testing (and thus the ability to adapt one's individual therapy) is only recommended in states $\big[\{m,h\}\,,0\,,0\,,0\big]$ in the resource-poor setting. It may therefore be indicated for resource-constrained settings, that despite the availability of subsidized treatment, their optimal use may not be feasible because informed decision making is not possible as a consequence of unaffordable diagnostics (k_{info} is too high). In the resource-rich setting (Germany) information costs produce 1.1% of the total costs, whereas they produce 3.3% of total costs in the resource-poor setting (South Africa); see table 4.2. Finally, it can be seen that the total expected costs (last rows in table 4.2) are disproportionately higher in Germany than in South Africa (i.e. compare their differences with the differences in GDP in table 3.2).

In order to quantify expected improvements when implementing the proposed policy, we compared the expected costs and probability of AIDS-related death with the standard-of-care treatment. Standard-of-care in Germany involves routine virological monitoring every 6 months. Upon detection of a viral rebound (any copy number state $n_C(M)=h$ or two consecutive $n_C(M)=m$), treatment is changed. In the resource-poor setting (South Africa), drug resistance tests are currently not part of the standard-of-care. Instead, treatment is changed if symptoms of immunodeficiency appear while the patient is under treatment with first line therapy a_1. We assumed that symptoms appear 6 months after a viral rebound. The predicted costs from standard-of-care therapies are shown in table 4.2, indicating that an overall improvement of the optimal policy vs. standard-of-care therapy is achieved, in particular in terms of state costs $V_{\mathcal{S}}(x)$.

The differences in diagnostic testing as well as the different treatment policies in the presence of highly resistant strains motivated us to assess whether cost-optimal policies may yield better outcomes in terms of patient life-expectancy if a) the costs for diagnostic tests k_{info} are reduced and b) if the costs of treatment are reduced.

4.2. Parameter sensitivity with regard to survival. The probability of AIDS-related death was computed using a well-known Monte-Carlo method [21].

		k_{info}					
		s-o-c§	basel.*	200	100	50	5
Germany	$V_{\text{info}}(x)$	7,816	10,502	7,681	5,393	3,684	950
	$V_{\mathcal{A}}(x)$	195,534	208,100	211,150	212,950	214,030	215,510
	$V_{\mathcal{S}}(x)$	794,857	705,380	698,180	693,710	690,900	686,880
	$\sum$	998,207	923,982	917,011	912,053	908,624	903,340
S. Africa	$V_{\text{info}}(x)$	0	2,294	2,467	1,838	1,314	369
	$V_{\mathcal{A}}(x)$	1,880	1,739	1,899	1,965	1,909	1,842
	$V_{\mathcal{S}}(x)$	74,014	65,116	62,408	60,928	60,928	60,104
	$\sum$	75,894	69,149	66,774	64,731	64,151	62,315

TABLE 4.2. ***Cost splitting.*** *Calculated cost splitting (in US\$) for state $x = [h\,,\,0\,,\,0\,,\,0]$ in the resource-rich and resource-poor settings, respectively. § 's-o-c' denotes standard-of-care treatment. ⋆ Baseline costs for resistance tests were 400 and 500 US\$ in Germany and South Africa, respectively.*

In the standard-of-care simulations, the probability of AIDS-related death for the resource-rich setting after 1000-, 3000-, and 5000 days of treatment were 11.7%, 28.9%, and 42.1% (see also figure 4.1A, open cyan circles), and 12.0%, 28.4%, and 41.7% in the resource-poor setting, respectively (see figure 4.1C, open cyan circles).

When applying the optimal policy with baseline parameters to the HIV dynamics model, a survival improvement in comparison to standard-of-care was achieved. In the resource-rich setting the probability of AIDS-related death was 5.0%, 15.5%, and 25.2% after 1000-, 3000-, and 5000 days of treatment (see also figure 4.1A, blue circles). In the resource-poor setting, the risk of AIDS-related death was 5.1%, 16.3%, and 31.0%, respectively (see figure 4.1C, blue circles). The slightly higher death probability under an optimal treatment policy in the resource-poor setting, in comparison to the resource-rich setting, may be a result of the inability to change treatment in time ($\tau \geq \tau_{\max}$ for many states in table 4.1). In the sequel, we evaluated whether reduced costs for diagnostic test may further improve survival. In figure 4.1A&C, we show the probability of AIDS-related death for reduced prices of drug resistance tests $k_{\text{info}} = 200, 100, 50, 5$ US\$ (green squares, red downward-pointing triangles, black upward-pointing triangles, and magenta diamonds). It can be seen that a reduction in diagnostic test prices may significantly improve patient survival in resource-poor settings and that the difference becomes more evident if later time points are evaluated (panel C). For resource-rich countries, patient survival is only insignificantly altered (panel A). To visualize the benefit of reduced diagnostic test prices, we show the 5000 days probability of AIDS-related death as a function of the price reduction factor for drug resistance tests in figure 4.1B & D. It can be seen that a price reduction of factor 2.5 (200 US\$ per test) in the resource-poor setting may already enable a level of death prevention similar to the resource-rich setting. In the resource-poor setting (panel D) the probability of AIDS-related deaths 13.7 years (5000 days) after treatment initiation were 31%, 24.2%, 21.8%, 19.2%, and 17.6% for test costs $k_{\text{info}} = 500, 200, 100, 50, 5$ US\$ per test, respectively. The probability of AIDS-related deaths 13.7 years (5000 days) after treatment initiation in the resource-rich setting were 25.2%, 24.1%, 22.7%, 21.4%, and 20.1% for test costs $k_{\text{info}} = 400, 200, 100, 50, 5$ US\$ per test, respectively.

Despite the (anticipated) reduction of information costs V_{info}, table 4.2 also reveals that the costs for the states $V_{\mathcal{S}}$ are reduced in the two settings (indicating a treatment

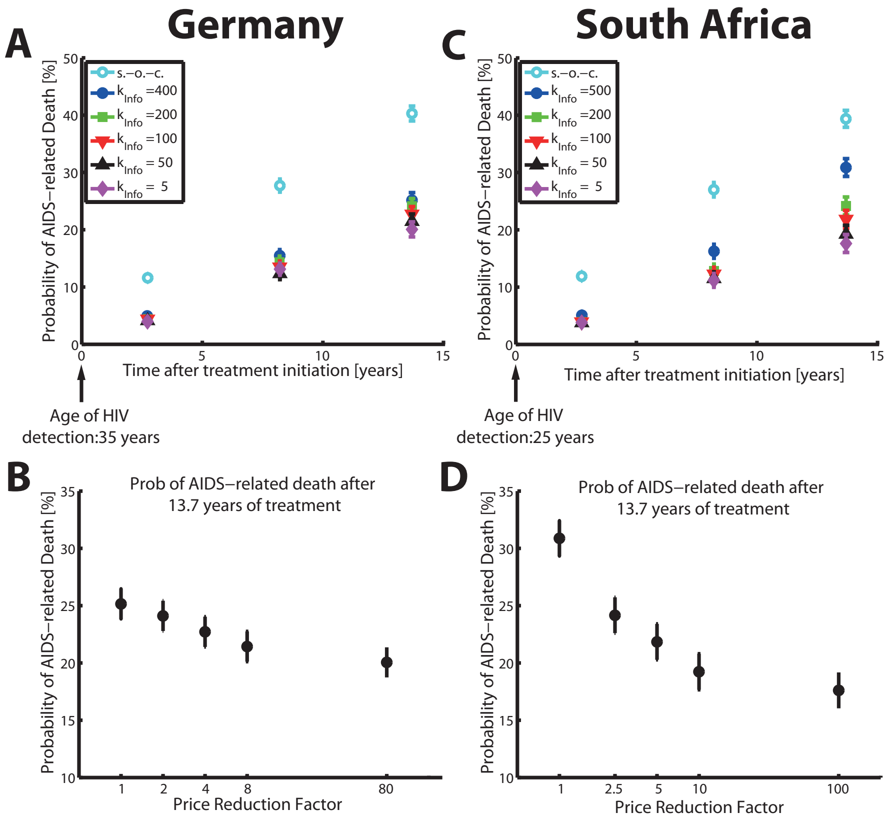

FIG. 4.1. *Effect of resistance test k_{info} cost reduction on AIDS survival. A: Probability of AIDS-related death 1000-, 3000-, and 5000 days after treatment initiation under application of* **a standard-of-care treatment** *(open cyan circles) vs. a cost-optimal policy with drug resistance test costs $k_{\text{info}} = 400, 200, 100, 50, 5$ US$ per test (blue dots, green squares, red downward-pointing triangles, black upward-pointing triangles, and magenta diamonds) in a resource-rich setting (Germany). B: Probability of AIDS-related death 5000 days after treatment initiation as a function of resistance test price reduction in a resource-rich setting (Germany). C: Probability of AIDS-related death 1000-, 3000-, and 5000 days after treatment initiation under application of* **a standard-of-care treatment (open cyan circles)** *vs. a cost-optimal policy with drug resistance test costs $k_{\text{info}} = 500, 200, 100, 50, 5$ US$ per test (blue dots, green squares, red downward-pointing triangles, black upward-pointing triangles, and magenta diamonds) in a resource-constrained setting (South Africa). D: Probability of AIDS-related death 5000 days after treatment initiation as a function of resistance test price reduction in a resource-constrained setting (South Africa).*

benefit). For Germany $V_{\mathcal{S}}$ is reduced by 2.6% when comparing the baseline parameters with $k_{\text{info}} = 5$ US$ per test, whereas it is 7.7% for South Africa. Interestingly, the cost of information (V_{info} in table 4.2) is not reduced for South Africa, when the price for diagnostics is reduced from $k_{\text{info}} = 500$ to $k_{\text{info}} = 200$ US$, while at the same time the total costs are reduced, which can be fully attributed to a state cost reduction $V_{\mathcal{S}}$. This indicates that the price reduction for diagnostics enables their more frequent use (thus no V_{info} reduction), which seems to fully benefit the patient (lower $V_{\mathcal{S}}$).

For the resource-rich setting, we also evaluated whether treatment cost reduction would improve patient survival. We found that there was only a small effect of treatment cost reduction on patient survival. The probability of AIDS-related death 13.7

years (5000 days) after treatment initiation were 23.8%, 23.5%, 23.6%, and 21.3% if treatment cost were reduced 2-, 4-, 10-, and 20-fold, respectively.

In conclusion, it may be said that prices for diagnostic test costs are too high in resource-poor settings. A small price reduction on the other hand may significantly improve patient survival in a cost-optimal way. For the resource-rich setting it is not indicated that price-reduction for either diagnostics or treatment would further improve patient survival in a cost-optimal way. The simultaneous price reduction of treatment and diagnostics, however, may do so.

5. Discussion & conclusions

We have presented a novel Markov Control Model which requires full observability of the Markov Process at certain rare observation times t_j, whereas the process is unobserved between two of these observation times. The process is therefore different from a Partially Observable Markov Decision Processes (POMDP), which usually assumes that the state cannot be determined explicitly and which takes a propagation of this uncertainty into account.

Moreover, we assume that each observation produces costs, precluding frequent testing. Therefore, one central challenge is to determine the (rare) observation times. The problem of finding an optimal balance between *information* and *cost reduction* is a new aspect regarding Markov Control Theory.

Our approach is motivated by the fact that in many applications (in particular medical), continuous observation and interaction is rarely feasible. Likewise, information is usually not freely available in real world situations.

Of course, the exactness of information is not self-evident either. However, in the presented application (as well as in many other applications), sufficient accuracy of the measurements can be justified by the coarse-grained representation of the underlying dynamics (the simplified HIV model in figure 3.1).

We applied the novel method to determine optimal HIV treatment switching and monitoring policies. We found distinct cost-optimal treatment policies for the case when two treatment lines a_1 and a_2 are available in (i) resource-rich (Germany) and (ii) resource-poor (South Africa) settings. For the resource-rich setting, we found that diagnostic testing should be performed more frequently in the case where high virus copy numbers are observed and less frequently when low numbers are encountered. Note, that currently virologic testing is recommended every 6 months, irrespective of the number of viruses present. Secondly, for resource-poor countries, we found that drug resistance testing may not be cost-optimal in most disease states when taking the current prices of diagnostics into account. Currently, drug resistance tests are not part of the standard-of-care in resource-poor settings. Although the computed optimal policy may already improve survival (see figure 4.1C), a small subsidy for resistance tests (2.5 fold reduction) may further improve patient survival by enabling informed decision making.

The presented work also has some limitations in application to real-world examples: In resource-rich countries like Germany, more sophisticated treatment changes may be implemented. Thus, more than two treatment options may be available. In the third world, however, this is not the case. Usually, a certain first-line treatment (a_1) and a second line treatment (a_2) are available (as in our application).

For logistical reasons, diagnostic assessment may be performed in a cross-sectional way in resource-constrained settings, i.e. all patients in a district are initiated and monitored at the same time. For this type of treatment practice, the herein derived approach may indicate when the initial assessments should occur. After a short while however, individualization as a result of the stochastic nature of the process would be required.

Although treatment has become available and affordable in resource-constrained setting through subsidy, clinical testing remains costly. Thus, the effective use of the available drugs may not be granted, as the prices for laboratory testing and thus informed decision making are too high to be mastered by the national economies. Our results support recent appeals for affordable drug resistance tests in the public sectors of resource-constrained settings [22]. We showed that price reductions for drug resistance tests may in fact improve patient survival in a cost-optimal way. Furthermore, sub-optimal treatment scheduling drives the emergence of drug resistance, which is a major problem in long-term epidemiologic control and could potentially be avoided if clinical testing was subsidized similarly to the treatment itself.

In the complete absence of drug resistance tools to guide informed decision making in resource-constrained settings, it may be beneficial to implement population-based treatment switching policies, i.e. treating all individuals for a certain period with treatment a_1 and than to switch to a_2, without having detailed knowledge about the actual state of the patient, e.g. [12]. Mathematically, the optimal (population-based) switching rule could be computed by interpreting the system as a "switched system", where the probability flux is deterministically steered according to the infinitesimal generators of respective treatments. This work will, however, be left for the future.

Appendix A. Bellman equation: proof. Introduce, for a given policy $u : S \to \mathcal{A} \times [0,\infty)$, the short notations $X_j := X_{t_j}$, $a_j := a(X_{t_j})$, $\tau_j := \tau(X_{t_j})$, $j = 0, 1, 2, ...$ I.e., we consider the discrete time stochastic processes $(X_j)_{j=0,1,2,...}$, $(a_j)_{j=0,1,2,...}$, $(\tau_j)_{j=0,1,2,...}$ of states, actions, and time intervals determined at each observation point. Note that these processes highly depend on each other. We make the following observations.

(i) By the short notation we can write

$$J(x,u) = \mathbb{E}_x^u\left(\sum_{j=0}^{\infty} e^{-\lambda t_j}\Big(C(X_j,a_j,\tau_j) + e^{-\lambda\tau_j}k_{\text{info}}\Big)\right).$$

(ii) For any function $V : S \to \mathbb{R}$ and any policy $u : S \to \mathcal{A} \times [0,\infty)$ it holds that

$$V(x) - \mathbb{E}_x^u\Big(e^{-\lambda t_{n+1}}T_{a_n,\tau_n}V(X_n)\Big) = \mathbb{E}_x^u\left(\sum_{j=0}^{n} e^{-\lambda t_j}\left(V(X_j) - e^{-\lambda\tau_j}T_{a_j,\tau_j}V(X_j)\right)\right).$$

To see this, we start from the right:

$$\begin{aligned}
&\mathbb{E}_x^u\left(\sum_{j=0}^{n} e^{-\lambda t_j}\left(V(X_j) - e^{-\lambda\tau_j}T_{a_j,\tau_j}V(X_j)\right)\right)\\
&\overset{(a)}{=} \sum_{j=0}^{n}\Big(\mathbb{E}_x^u\Big(e^{-\lambda t_j}V(X_j)\Big) - \mathbb{E}_x^u\left(e^{-\lambda t_j}e^{-\lambda\tau_j}T_{a_j,\tau_j}V(X_j)\right)\Big)\\
&\overset{(b)}{=} \sum_{j=0}^{n}\Big(\mathbb{E}_x^u\Big(e^{-\lambda t_j}V(X_j)\Big) - \mathbb{E}_x^u\Big(e^{-\lambda t_{j+1}}\mathbb{E}^u\Big(V(X_{j+1})|X_j\Big)\Big)\Big)
\end{aligned}$$

$$
\begin{aligned}
&\overset{(c)}{=} V(x) + \sum_{j=0}^{n-1} \left(\mathbb{E}_x^u\Big(e^{-\lambda t_{j+1}} V(X_{j+1})\Big) - \mathbb{E}_x^u\Big(e^{-\lambda t_{j+1}} \mathbb{E}^u\big(V(X_{j+1})|X_j\big)\Big)\right) \\
&\qquad - \mathbb{E}_x^u\Big(e^{-\lambda t_{n+1}} \mathbb{E}^u\big(V(X_{n+1}|X_n)\big)\Big) \\
&\overset{(d)}{=} V(x) - \mathbb{E}_x^u\Big(e^{-\lambda t_{n+1}} T_{a_n,\tau_n} V(X_n)\Big).
\end{aligned}
$$

In (a) we used the linearity of the expected value. (b) follows from the fact that $T_{a_j,\tau_j}V(X_j) = \mathbb{E}^u(V(X_{j+1})|X_j)$ and $e^{-\lambda t_j}e^{-\lambda \tau_j} = e^{-\lambda t_{j+1}}$ by definition. Step (c) is a rearrangement of the sum and the replacement $\mathbb{E}_x^u\big(e^{-\lambda t_0}V(X_0)\big)= V(x)$ (using $t_0 = 0$ and $X_0 = x$). In (d) the whole sum disappears due to the law of total expectation in the sense of ted value. (b) follows from the fact that $T_{a_j,\tau_j}V(X_j) = \mathbb{E}^u(V(X_{j+1})|X_j)$ and $e^{-\lambda t_j}e^{-\lambda \tau_j} = e^{-\lambda t_{j+1}}$ by definition. Step (c) is a rearrangement of the sum and the replacement $\mathbb{E}_x^u\big(e^{-\lambda t_0}V(X_0)\big)= V(x)$ (using $t_0 = 0$ and $X_0 = x$). In (d) the whole sum disappears due to the law of total expectation in the sense of

$$
\begin{aligned}
\mathbb{E}_x^u\Big(e^{-\lambda t_{j+1}} V(X_{j+1})\Big) &= \mathbb{E}_x^u\Big(\mathbb{E}\big(e^{-\lambda t_{j+1}} V(X_{j+1})|X_j, t_j\big)\Big) \\
= \mathbb{E}_x^u\Big(e^{-\lambda t_{j+1}} \mathbb{E}\big(V(X_{j+1})|X_j, t_j\big)\Big) &= \mathbb{E}_x^u\Big(e^{-\lambda t_{j+1}} \mathbb{E}\big(V(X_{j+1})|X_j\big)\Big).
\end{aligned}
$$

Here we used first that, given X_j and t_j, t_{j+1} is a deterministic quantity such that $e^{-\lambda t_{j+1}}$ can be taken out of the expected value like a scalar factor, and second that X_{j+1} only depends on X_j but not on t_j as our dynamics are assumed to be time homogeneous. ted value like a scalar factor, and second that X_{j+1} only depends on X_j but not on t_j as our dynamics are assumed to be time homogeneous.

(iii) If a function V satisfies the given Bellman equation, it fulfills

$$
V(X) \le C(x, a, \tau) + e^{-\lambda\tau} k_{\text{info}} + e^{-\lambda\tau} T_{a,\tau} V(x)
$$

for any $a \in \mathcal{A}, \tau \ge 0$. That is,

$$
V(X) - e^{-\lambda\tau} T_{a,\tau} V(x) \le C(x, (a, \tau)) + e^{-\lambda\tau} k_{\text{info}}.
$$

Putting (ii) and (iii) together one gets

$$
V(x) - \mathbb{E}_x^u\Big(e^{-\lambda t_{n+1}} T_{a_n,\tau_n} V(X_n)\Big) \le \mathbb{E}_x^u\left(\sum_{j=0}^{n} e^{-\lambda t_j}\Big(C(X_j, a_j, \tau_j) + e^{-\lambda\tau_j} k_{\text{info}}\Big)\right).
$$

Now taking $n \to \infty$ we get

$$
V(x) \le J(x, u),
$$

where u was any given policy. In this last step we used on the left hand side the fact that $T_{a_n,\tau_n}V(X_n)$ is bounded while $e^{-\lambda t_{n+1}}$ converges to zero (as $t_{n+1} \to \infty$), and on the right hand side we used monotone convergence and observation (i).
We obtain equality when choosing $u = u^*$.

Appendix B. Parameter estimation for the HIV-model. We estimated all parameters p by fitting the presented model to clinical data of virus decay- and rebound dynamics [23, 24, 25, 26] as well as clinically observed probabilities of virologic suppression during first-line a_1 and second-line a_2 therapy in treatment-naive and treatment-experienced patients [27, 28, 29, 30] by minimizing the following weighted least-square criterion:

$$\arg\min_p \sum_i \sum_j \sum_{x\in\mathcal{S}} \left(\left(\frac{\tilde{P}_i\big(x, t_j|\pi_0, t_0, p\big) - P_i\big(x, t_j|\pi_0, t_0\big)}{P_i\big(x, t_j|\pi_0, t_0\big)} \right) \cdot N_{ij} \right)^2,$$

where $p = \Big(\delta_h, \delta_m, k_{\ell,\emptyset}, k_{m,\emptyset}, \mu_{h,\mathrm{R1},\emptyset}, \mu_{m,\mathrm{R1},\emptyset}, \mu_{h,\mathrm{R2},\emptyset}, \mu_{m,\mathrm{R2},\emptyset}, \eta(a_1), \eta(a_2)\Big)$ denotes the set of parameters to be estimated and $\tilde{P}_i\big(x, t_j|\pi_0, t_0, p\big)$ denotes the model predicted probability that the population vector is x at time t_j for clinical/experimental condition i, given parameters p. $P_i\big(x, t_j|\pi_0, t_0\big)$ denotes the corresponding clinically observed probability (in the sense of frequency). The initial distribution π_0 at the time of treatment initiation t_0 were either estimated from the data [23, 24, 25, 26] or set to $\pi_0(\big[h\,,\,0\,,\,0\,,\,0\big]) = 1$ in treatment-naive patients and to $\pi_0(\big[\ell\,,\,h\,,\,0\,,\,0\big]) = 1$ in treatment-experienced patients. The parameter N_{ij} denotes the number of individuals that gave rise to the clinically observed probabilities $P_i\big(x, t_j|\pi_0, t_0\big)$ for clinical condition i at time t_j. We constrained the parameter space to non-negative numbers and bounded the values for the efficacy of the drugs to $0 \leq \eta(a_1) \leq 1$ and $0 \leq \eta(a_2) \leq 1$, respectively. The following inequalities had to be fulfilled: (i) $k_{\ell,\emptyset} \leq k_{m,\emptyset}$ and (ii) $\delta_m \leq k_{\ell,\emptyset}$, which is motivated by population dependent virus growth (inequality (i)) and the fact that the virus should rather grow than decline in the absence of treatment (inequality (ii)).

The agreement between the model-predicted probabilities $\tilde{P}_i\big(x, t_j|\pi_0, t_0, p\big)$ and clinically observed probabilities $P_i\big(x, t_j|\pi_0, t_0\big)$ was generally good, although we observed deviations in the short-term dynamics of viral decay and rebound that can be attributed to the coarse-graining of the underlying viral kinetics (its representation in terms of only four copy number states $0, \ell, m, h$). The agreement with experimental data could, however, be further improved if more copy number states were utilized (less coarse graining of the underlying dynamics).

Acknowledgments. MvK acknowledges support from the DFG research center MATHEON through project A21 “Modeling, Simulation and Therapy Optimization for Infectious Diseases”.

REFERENCES

[1] G.A. Pavliotis, *Stochastic Processes and Application*, (lecture notes; available at `http://http://www2.imperial.ac.uk/~pavl/`, accessed 06-august-2012), 2010.

[2] D.J. White, *A survey of applications of Markov decision processes*, J. Opl. Res. Soc., 44, 1073–1096, 1993.

[3] R. Luo, M.J. Piovoso, J. Martinez-Picado, and R. Zurakowski, *Optimal antiviral switching to minimize resistance risk in HIV therapy*, PLoS One, 6:e27047, 2011.

[4] E.A. Hernandez-Vargas, R.H. Middleton, and P. Colaneri, *Optimal and MPC switching policies for mitigating viral mutation and escape*, IFAC World Congress Milano (Italy), 28(2), 14857–14862, 2011.

[5] I.M. Rouzine, A. Rodrigo, and J.M. Coffin, *Transition between stochastic evolution and deterministic evolution in the presence of selection: General theory and application to virology*, Microbiol Mol. Biol. Rev., 65, 151–185, 2001.

[6] S.M. Shechter, M.D. Bailey, and A.J. Schaefer, *A modeling framework for replacing medical therapies*, IIE Transactions, 40, 861–869, 2008.

[7] X. Guo and O. Hernandez-Lerma, *Continuous-time Markov Decision Processes: Theory and Applications*, Stochastic Modelling and Applied Probability, Springer, Heidelberg, 2009.

[8] W.H. Fleming and R.W. Rishel, *Deterministic and Stochastic Optimal Control*, Appl. Math., Springer, Heidelberg, 1975.

[9] R.A. Howard, *Dynamic Programming and Markov Processes*, MIT Press, 1960.

[10] Y. Ye, *The simplex and policy-iteration methods are strongly polynomial for the Markov decision problem with fixed discount rate*, Math. Op. Res., 36, 593–603, 2011.

[11] M. von Kleist, S. Menz, and W. Huisinga, *Drug-class specific impact of antivirals on the reproductive capacity of HIV*, PLoS Comp. Bio., 6:e1000720, 2010.

[12] M. von Kleist, S. Menz, H. Stocker, K. Arasteh, Ch. Schütte, and et al., *HIV quasispecies dynamics during pro-active treatment switching: impact on multi-drug resistance and resistance archiving in latent reservoirs*, PLoS One, 6, e18204, 2011.

[13] D. Finzi, J. Blankson, J.D. Siliciano, J.B. Margolick, K. Chadwick, and et al., *Latent infection of CD4+ T cells provides a mechanism for lifelong persistence of HIV-1, even in patients on effective combination therapy*, Nat. Med., 5, 512–517, 1999.

[14] O. Lambotte, M.L. Chaix, B. Gubler, N. Nasreddine, C. Wallon, and et al., *The lymphocyte HIV reservoir in patients on long-term HAART is a memory of virus evolution*, AIDS, 18, 1147–1158, 2004.

[15] J. Coffin, F. Maldarelli, S. Palmer, A. Weigand, S. Brun, and et al., *Long-term persistance of low-level HIV-1 in patients on suppressive antiretroviral therapy*, abstract 169. 13th Conference on Retroviruses and Opportunistic Infections; 5–8 February 2006; Denver, Colorado, United States, 2006.

[16] P. Sendi, H.F. Günthard, M. Simcock, B. Ledergerber, J. Schüpbach, and et al., *Cost-effectiveness of genotypic antiretroviral resistance testing in HIV-infected patients with treatment failure*, PLoS One, 2:e173, 2007.

[17] UNAIDS, Report on the Global AIDS Epidemic 2010 (available at: `http://www.unaids.org/documents/20101123_globalreport_em.pdf`, accessed 25-May-2012).

[18] M. Stoll, C. Kollan, F. Bergmann, J. Bogner, and G. Faetkenheuer, *Calculation of direct antiretroviral treatment costs and potential cost savings by using generics in the German HIV ClinSurv cohort*, PLoS One, 6:e23946, 2011.

[19] The Clinton Health Access Initiative. Antiretroviral (ARV) Ceiling Price List (available at `http://www.clintonfoundation.org`, accessed 25-May-2012), 2011.

[20] The International Monetary Fund. World economic outlook database (available at `http://www.imf.org/external/pubs/ft/weo/2012/01/weodata/index.aspx`, accessed 25-May-2012).

[21] D.T. Gillespie, *A general method for numerically simulating the stochastic time evolution of coupled chemical reactions*, J. Comput. Phys., 22, 403–434, 1976.

[22] BioAfrica, Report of the 4th South African HIV Drug Resistance and Treatment Monitoring Workshop (available at: `http://bioafrica.mrc.ac.za/workshops/PDFs/Report4thSouthAfrican.pdf`, accessed 25-May-2012), 2009.

[23] M. Markowitz, M. Louie, A. Hurley, E. Sun, and M. Di Mascio, *A novel antiviral intervention results in more accurate assessment of human immunodeficiency virus type 1 replication dynamics and T-cell decay in vivo*, J. Virol., 77, 5037–5038, 2003.

[24] A.S. Perelson, P. Essunger, Y. Cao, M. Vesanen, and A. Hurley, *Decay characteristics of HIV-1-infected compartments during combination therapy*, Nature, 387, 188–191, 1997.

[25] L. Ruiz, J. Martinez-Picado, J. Romeu, R. Paredes, and M.K. Zayat, *Structured treatment interruption in chronically HIV-1 infected patients after long-term viral suppression*, AIDS, 14, 397–403, 2000.

[26] P.R. Harrigan, M. Whaley, and J.S. Montaner, *Rate of HIV-1 RNA rebound upon stopping antiretroviral therapy*, AIDS, 13, F59–F62, 1999.

[27] S. Staszewski, J. Morales-Ramirez, K.T. Tashima, A. Rachlis, and D. Skiest, *Efavirenz plus zidovudine and lamivudine, efavirenz plus indinavir, and indinavir plus zidovudine and lamivudine in the treatment of HIV-1 infection in adults*, N. Engl. J. Med., 341, 1865–1873, 1999.

[28] J.R. Arribas, A.L. Pozniak, J.E. Gallant, E. Dejesus, and B. Gazzard, *Tenofovir disoproxil fumarate, emtricitabine, and efavirenz compared with zidovudine/lamivudine and efavirenz in treatment-naive patients: 144-week analysis*, J. Acquir. Immune. Defic. Syndr., 47, 74–78, 2008.

[29] J.K. Rockstroh, J.L. Lennox, E. Dejesus, M.S. Saag, and A. Lazzarin, *Long-term treatment with raltegravir or efavirenz combined with tenofovir/emtricitabine for treatment-naive human immunodeficiency virus-1-infected patients: 156-week results from STARTMRK*, Clin. Infect. Dis., 53, 807–816, 2011.

[30] R.T. Steigbigel, D.A. Cooper, P.N. Kumar, J.E. Eron, and M. Schechter, *Raltegravir with optimized background therapy for resistant HIV-1 infection*, N. Engl. J. Med., 359, 339–354, 2008.