Many organizations strive to attain a balance between expanding their services to new customers and maintaining quality of service for existing customers. This trade-off becomes particularly acute when the organization faces uncertainty in the availability of a key resource. In this paper, we study this trade-off in a specific operational context, that of scaling up HIV treatment programs in sub-Saharan Africa. HIV clinics in this context receive extremely limited and uncertain supply of antiretroviral drugs (ARVs) that needs to be used for initiating untreated patients on treatment and for continuing treatment for patients who have been previously treated. On the one hand, clinics can focus on ensuring continuous treatment to their previously treated patients [Schouten et al., 2011] by being conservative in their scale-up and capping the number of new enrollments. However, this can lead to delay in treatment initiation for untreated patients, which in turn can lead to significant reduction in their quality of life and even death due to disease progression. On the other hand, clinics can be aggressive by enrolling many new patients but increase the risk of their treatment interruption in future periods due to drug stock-outs, which in turn can lead to adverse clinical outcomes such as treatment failure [Hamers et al., 2012], drug resistance [WHO et al., 2013] and increased mortality [El-Sadr et al., 2006], and which can necessitate transitioning patients to a much more expensive second line of therapy.

We model this trade-off embedded in the scale-up of ARV treatment programs using a stochastic dynamic programming framework. We classify HIV patients from a clinic’s catchment area into the following categories: (i) patients who are not yet clinically eligible to be initiated on treatment based on national guidelines (ineligible), (ii) patients who are clinically eligible but have not yet been initiated on treatment (eligible and untreated), (iii) patients who have been initiated on treatment earlier and are still responsive to it (eligible and treated), and (iv)
patients who had been previously initiated on treatment but are now resistant to it due to previous treatment interruptions (resistant). Each period, based on the inventory of ARVs and the number of patients in each of the above four categories, the clinic decides the number of treated patients and untreated patients to treat and sets aside any remaining inventory to be carried over to the next period. Then the following transitions occur: (i) new infections are added to the category of ineligible patients (ii) a fraction of ineligible patients move to the category of eligible patients due to natural progression of their disease (iii) some of the untreated patients move to the category of treated patients upon enrollment and (iv) a fraction of treated patients whose treatment is interrupted in this period move to the category of resistant patients. Following these transitions, a fraction of patients in each category die and the surviving patients obtain a reward, which is equal to the per-period quality of life utility associated with their category. At the end of the period, the clinic receives a shipment of ARVs of an uncertain quantity. The objective of the clinic is to maximize the total expected quality adjusted life years of its patients over the planning horizon subject to availability of drugs and the number of patients in different segments.

Analytical difficulties preclude complete characterization of the optimal policy for the most general version of the problem. Consequently, we provide a partial characterization of the optimal policy for certain special cases, which consists of dynamic prioritization of patient segments characterized by state-dependent thresholds and switching curves. Importantly, we find that it is optimal to not fully serve the segment of previously treated patients in certain cases despite having adequate inventory. This result is different from the traditional literature on multi-product and/or multi-customer inventory rationing models [Evans, 1967], where the optimal policy typically involves complete static prioritization of the high value segment over the low value segment. We show that the main driver for this structural difference is the endogenous movement of patients across segments depending on whether they received the product in the current period; an effect that is not considered in the literature. However, this feature also presents an ethical dilemma and makes the optimal policy an unlikely candidate for implementation in addition to its complex structure.
Hence, we develop a Two-Period heuristic by combining insights from a simpler two-period optimal policy and those from special cases of the n-period optimal policy. This heuristic is amenable to implementation and ensures that all previously treated patients will be treated in any period as long as there is adequate inventory in that period. Extensive computational experiments show that the average optimality gap of this heuristic is less than 4% over a wide range of parameter values. In contrast, a Safety-Stock heuristic based on current practice [Schouten et al., 2011] yields average optimality gaps of around 20%. Beyond better overall performance, the Two-Period heuristic is also much more robust to misspecification of the parameter values and is no more difficult to implement than the Safety-Stock heuristic. We believe that the insights generated from our analysis can be used to develop a decision support tool for clinics to design their ARV treatment program scale-up plans.

References


