2004s-13

Early Initiation of Highly Active Antiretroviral Therapies (HAART) for HIV/Aids: The Contribution of a Stochastic Dynamic Model of Choice

Pierre Lasserre, Jean-Paul Moatti, Antoine Soubeyran

Série Scientifique
Scientific Series

Montréal
Mars 2004

CIRANO
Centre interuniversitaire de recherche
en analyse des organisations

CIRANO

Le CIRANO est un organisme sans but lucratif constitué en vertu de la Loi des compagnies du Québec. Le financement de son infrastructure et de ses activités de recherche provient des cotisations de ses organisations-membres, d'une subvention d'infrastructure du ministère de la Recherche, de la Science et de la Technologie, de même que des subventions et mandats obtenus par ses équipes de recherche.

CIRANO is a private non-profit organization incorporated under the Québec Companies Act. Its infrastructure and research activities are funded through fees paid by member organizations, an infrastructure grant from the Ministère de la Recherche, de la Science et de la Technologie, and grants and research mandates obtained by its research teams.

Les organisations-partenaires / The Partner Organizations

PARTENAIRE MAJEUR
. Ministère du développement économique et régional [MDER]

PARTENAIRES
. Alcan inc.
. Axa Canada
. Banque du Canada
. Banque Laurentienne du Canada
. Banque Nationale du Canada
. Banque Royale du Canada
. Bell Canada
. BMO Groupe Financier
. Bombardier
. Bourse de Montréal
. Caisse de dépôt et placement du Québec
. Développement des ressources humaines Canada [DRHC]
. Fédération des caisses Desjardins du Québec
. GazMétro
. Hydro-Québec
. Industrie Canada
. Ministère des Finances [MF]
. Pratt & Whitney Canada Inc.
. Raymond Chabot Grant Thornton
. Ville de Montréal

. École Polytechnique de Montréal
. HEC Montréal
. Université Concordia
. Université de Montréal
. Université du Québec à Montréal
. Université Laval
. Université McGill

ASSOCIE A :
. Institut de Finance Mathématique de Montréal (IFM^2)
. Laboratoires universitaires Bell Canada
. Réseau de calcul et de modélisation mathématique [RCM^2]
. Réseau de centres d'excellence MITACS (Les mathématiques des technologies de l'information et des systèmes complexes)

Les cahiers de la série scientifique (CS) visent à rendre accessibles des résultats de recherche effectuée au CIRANO afin de susciter échanges et commentaires. Ces cahiers sont écrits dans le style des publications scientifiques. Les idées et les opinions émises sont sous l'unique responsabilité des auteurs et ne représentent pas nécessairement les positions du CIRANO ou de ses partenaires.

This paper presents research carried out at CIRANO and aims at encouraging discussion and comment. The observations and viewpoints expressed are the sole responsibility of the authors. They do not necessarily represent positions of CIRANO or its partners.

ISSN 1198-8177

Early Initiation of Highly Active Antiretroviral Therapies (HAART) for HIV/aids: The Contribution of a Stochastic Dynamic Model of Choice[*]

Pierre Lasserre[†], *Jean-Paul Moatti*[‡], *Antoine Soubeyran* [§]

Résumé / *Abstract*

Le bien-fondé d'administrer précocement des traitements antirétroviraux à haute activité (HAART) aux personnes infectées par le VIH reste l'objet de débats dans le monde car leurs bienfaits à court terme peuvent compromettre les traitements futurs si des souches résistantes du virus se développent. Nous formulons un modèle qui combine irréversibilité et inertie dans un cadre de décision thérapeutique séquentielle. L'information se révèle entre la première et la deuxième période, si bien que la décision thérapeutique de deuxième période est conditionnelle à cette information. Il s'avère que des patients identiques peuvent se voir administrer des traitements différents à l'optimum; de plus, pour des paramètres justifiant des décisions bien tranchées en période 2 (à patients identiques traitement identique), la décision de période 1 est plus complexe car il est alors trop tard pour en renverser les conséquences en période 2. Autre résultat: supposons que le risque de résistance est élevé en cas d'échec thérapeutique du traitement initial (pays en développement; groupes défavorisés); nous montrons alors que les différences dans l'estimation de ce risque n'interviennent pas dans le choix optimal de la taille de la population qui se voit administrer le traitement antirétroviral. L'introduction du traitement relève alors purement de considérations d'efficacité et de coût.

Mots clés : décisions thérapeutiques, incertitude,information, irréversibilité, traitement, apprentissage, erreurs.

[*] Financial support from the FCAR funds, IFM2 and the CNRS is gratefully acknowledged. We thank two anonymous referees; one of them gave us comments too helpful and insightful to be given credit properly within the paper.

[†] Département des sciences économiques, Université du Québec à Montréal, GREQAM, and CIRANO.

[‡] Faculté d'économie et gestion, Université de la Méditerranée, GREQAM, and INSERM Research Unit 379.

[§] Département des sciences économiques, Université de la Méditerranée, and GREQAM. Address all correspondence to : GREQAM, Rte des Milles, 13290 Les Milles, France. Tel. : +33 (0) 4 42 93 59 80; e-mail : soubey@univ-aix.fr.

Criteria for initiation of highly active antiretroviral treatments (HAART) in HIV-infected patients remain a matter of debate world-wide because short-term benefits have to be balanced with costs of these therapies, and restrictions placed on future treatment options if resistant viral strains develop. In order to take into account irreversibility and inertia effects associated with ex ante choices, we propose a simple stochastic dynamic model of sequential therapeutic choice with intermediary revelation of information, in which the efficiency gains from a new effective Therapy in second period are conditional on the results of the treatment in the previous period. We find that identical patients may be administered different treatments at the optimum; for parameters implying an all or nothing cut decision in period 2, a more forward looking decision rule is required in period 1 because there will be little space for adjustment to its consequences. Another finding is that as soon as risks of resistance due to therapeutic failure of initial treatments are significant, as perhaps in developing countries and in marginalized groups of developed countries, differences in the estimation of this risk should not influence the optimal decision about the size of the HIV-infected population eligible for early antiretroviral Therapy. The decision should then be based on pure efficiency/cost considerations.

Keywords: *therapeutic decisions, uncertainty, information, irreversibility, treatment, learning, errors.*

Code JEL: I18, C61.

Contents

1. Introduction

During the last seven years, the clinical care of HIV-infected people has been substantially influenced by the introduction of the Highly Active Antiretroviral Treatment (HAART). This treatment combines several therapies associating new HIV-specific protease inhibitors (PIs), and, more recently, non nucleoside reverse transcriptase inhibitors (NNRTIs), with previously existing antiretroviral drugs, the Nucleoside Reverse Transcriptase Inhibitors (NRTI). Short-term studies (Hogg et al., 1998a; Murphy et al., 2001) have clearly proved HAART therapies to be effective in decreasing viral replication and reducing morbidity and mortality among HIV-infected patients. However, major uncertainties remain about the optimal criteria for therapy initiation, as well as the dosages and specific HAART combinations that will ensure long-term efficacy (Gallant, 2000).

Current recommendations for initiation of HAART in patients infected with human immunodeficiency virus type 1 HIV are based on the combination of two biological markers, CD4 T-lymphocyte cell counts and plasma HIV RNA levels. The relative prognostic value of each marker following initiation of therapy has not been fully characterized. Earlier guidelines were heavily based on the principle of 'hit early, hit hard', although the long-term implications of this approach were unknown (Kyriakides and Guarino, 2001). Since then, the only clear international consensus is that HAART should be initiated before CD4 cell counts become lower than 200/microL because uniformly higher rates of disease progression to death and AIDS or death among patients starting ARV therapy have been observed in patients having access to HAART below this threshold (Hogg, 2001). On the other hand, there is no clear evidence whether delay in initiation of ARV therapy (ART) until this threshold of 200/microL may lead to a poorer viral load response for patients with human immunodeficiency virus (Phillips et al., 2001). Due to lack of evidence, country guidelines may significantly differ for recommendations about patients with CD41 lymphocytes between 200 and 350 cells/mL. Differences are even more pronounced for patients with CD41 lymphocytes > 350 cells/mL and when combining CD4 thresholds with those used for viral load (Rubio, 2002; Recommendations of the Panel on Clinical Practices for Treatment of HIV, 2002; Idemyor, 2002).

Because of increased awareness of the activity and toxicity of current drugs, the threshold for initiation of therapy has shifted to a later time in the course of HIV disease. However, the optimal time to initiate therapy remains imprecisely defined (Yeni et al., 2002; Delfraissy, 2002). A major issue has to do with the development of resistance, and the subsequent loss of drug activity, which may be caused by a variety of pharmacological and biological (Descamps et al., 2000) as well as behavioral factors (Paterson et al., 2000). One potential long term consequence is cross-resistance to alternative HIV treatments not yet prescribed (Deeks et al., 1997).

The prospect of resistance is a key parameter for clinical decision-making about initiation of HAART because it threatens to make individual patients unresponsive to their first line regimen and to reduce the effectiveness of switching to other available or future regimens in case of failure of previous HAART combinations. Indeed, virological treatment failure has been reported in circa 50% of unselected patients within one year initiation of a PI-containing regimen (van Heeswijk, 2001), and such failure is more frequent when patients had received previous ARV treatment (Le Moing et al., 2002). As is already the case with resistance to antibiotics (Laxminarayan, 2002), the emergence of HIV viral strains which have become resistant to existing antiretroviral drugs also raises the spectre of a public health threat in case of new infection from HIV viral strains having acquired resistance against current therapies (Hecht et al.,1998 ; Wainberg & Friedland, 1998). Among persons in North America and Europe who are newly infected with the human immunodeficiency virus HIV the prevalence of transmitted resistance to ARV drugs has been estimated at 1 to 11 percent. The proportion of new infections that involve drug-resistant virus is increasing in developed countries and initial HAART is more likely to fail in patients who are infected with drug-resistant virus (Little et al., 2002; Ristig et al., 2002). As resistance to antibiotics, transmission of HIV resistant strains to newly infected patients is essentially an externality.

In this paper, we do not focus on that externality facing the healthy population, although as we explain in the discussion of Section 5.3., it could easily be adapted to study that particular aspect of resistance. We rather focus on how the prospect of resistance should be taken into account for optimizing treatment choices from the point of view of already infected patients. When dealing with this issue, a different externality arises from the fact that direct individual treatment costs are typically sensitive to the number of individuals under therapy. From a technical point of view, the individual medical decision to initiate an antiretroviral combination therapy, associating drugs from the three major classes currently available, is similar to the optimal timing of the acquisition of a new technology, when the latter is still in development and there are irreversibilities and inertia associated with the decision. Clinicians, patients, as well as public health authorities must make choices on the basis of data that are not fully matured, under major uncertainties about their consequences on HIV-infected persons. The biological plausibility of early initiation of HAART needs to be weighted against the potential toxicity of the therapy and the restrictions it may place on later treatment options, especially if resistant viral strains appear and are widely spread. Our methodolgical contribution here is to propose a model of choice that takes into account the fact that a clinician who initiates prescription of HAART makes a costly irreversible decision while the relevant body of scientific information is still evolving, implying that mistakes are possible *ex post*, even if the right decision was made *ex ante* at the time of initiation.

The model formalizes and solves the therapeutic dilemma whether or not, and when, initiate a therapy, as well as the economic dilemma of weighting financial cost considerations against health benefits. The formal resolution of the problem for various medical and economic parameter configurations illustrates how the methodology deals with evolving scientific uncertainty and can be applied to various therapeutic situations. It also leads to introducing notions such as marginal efficiency costs ratios, resistance to therapy, alternative treatment strategies, and to describing their intricate roles in the solution. While each has a clear rationale, some results are surprising. For example, in some configurations it is preferable to keep flexibility for future therapeutic decisions;

in such cases, it turns out that early decisions are disconnected from future costs, not because the outcome is independent of such costs - it is not - but because the future therapeutic strategy equalizes net outcomes over states of nature. Also surprising is the possibility, due to the presence of non convexities, that it might be optimal to treat identical patients in unequal ways, or that the prospect of resistance to future therapy be more often irrelevant to current decisions, the higher that resistance effect.

In practice, the formulation of guidelines and criteria for initiating treatments, will determine the proportion of patients immediately benefiting from them but also the level of risk that resistant viral strains will emerge in an already infected person. This decision can also be considered equivalent to the optimal timing of an investment during the course of development of an innovation. Industrial economics has already been confronted with other, similar, problems such as quality choice or vertical differentiation. Although much of the literature in this field was static or allowed commitment (Stokey, 1989), it is now considered that the crux of the problem is indeed the timing of decisions, so that early commitment is a bad assumption. A stochastic dynamic programming approach is more appropriate in such circumstances (Ross, 1983). In particular, the real options literature typically deals with the arrival of new information after some costly, irreversible, decision has been taken.

The rest of the paper is organized as follows. In the next section we introduce a model that exhibits the main features just discussed qualitatively: irreversible therapeutic decisions are made under uncertainty in a dynamic setup; information unfolds during the period under scrutiny so that decisions may be regretted *ex post* even if they were not mistakes *ex ante*. The problem is solved by stochastic dynamic programming, starting, in Section 3, with the last period, and continuing, in Section 4, with the first period and the overall solution. Section 5 discusses various aspects of the problem and results, in particular the importance of costs and efficiency costs ratios, public good aspects and the issue of equal treatment of equals, the relevant alternative treatments available to the decision maker, the issue of resistance, and the link with real options. In the

Conclusion, we go back to the general medical literature and raise issues that could not be tackled without further work.

2. Sequential therapeutic choice with intermediary revelation of information

We use a simple two-period model to mimic the choices faced by clinicians confronted with a successive flow of new therapies with major *ex ante* uncertainties about their effectiveness. At the beginning of the first period, a new therapy, Therapy 1, becomes available. Although its effectiveness is not yet fully established, clinicians decide whether or not to prescribe it to some or all patients. The therapy has two effects: a known current utility improvement δ for the patient; and an unknown future resistance effect which will affect non only the utility of Therapy 1 in the future, but also the utility of a second therapy which will become available in the second period.

At the beginning of period 1, a fraction of the total number of HIV-infected patients are prescribed Therapy 1, assumed to be 'good' (g) with probability γ, and 'bad' (b) otherwise; the remaining fraction is not treated. Clinical experiments such as viral load measurements performed during period 1 then determine the resistance induced by Therapy 1: g or b.

At the beginning of the second period, the information about the effectiveness of Therapy 1 is revealed and, at the same time, a new therapy becomes available, with *ex-ante* probabilities of success or failure Γ and $1-\Gamma$. Again one may think of a successful Therapy 2 (G for good) as having a lasting beneficial effect, and the converse with a bad therapy (B). Since period 2 is the last period that we consider, Therapy 2 will be assumed to yield patients a higher utility improvement if it turns out good (G) than in the opposite case.[1]

Total treatment costs are a function of the total number of patients undergoing

[1] Unless otherwise mentioned, we use low-case letters for variables and functions pertaining to period 1, and capitals for variables and functions pertaining to period 2.

treatments 1 or 2. While Therapy 2 may well be more costly, taking this into account would complicate the notation without making any obvious differences. More interesting is the issue of economies of scale. It makes sense to take either a long-run perspective, with linear total costs, or a short-run view, with increasing unit costs. While the former may be well suited to discuss *ex ante* decision making, the short-run view corresponds to the situation of scare resources which clinicians usually face: increasing the number of patients undergoing therapy usually imposes a strain on fixed equipments and may require the acquisition of new fixed equipments that would otherwise not be required. Thus we hypothesize an increasing convex cost function of the number of patients z under therapy. For period 1 and period 2, total costs, discounted to period 1, are respectively

$$\begin{aligned} c(z) &= \frac{1}{2}cz^2 \text{ for period 1} \\ C(Z) &= \frac{1}{2}CZ^2 \text{ for period 2} \end{aligned}$$

where c and C are positive parameters that reflect technology and the constraints of health services, but also take account of discounting and the possibility that periods 1 and 2 might be of different durations.

Let t be the treatment undergone by a patient in period 1: no treatment, or Therapy 1; if Therapy 1 is applied $(t = t')$, it reveals itself either successful or unsuccessful:

$$t = \begin{cases} 0 \;\; \text{(no treatment)} \\ t' = \begin{cases} g \;\; \text{(treatment with Therapy 1, which is good)} \\ b \;\; \text{(treatment with Therapy 1, which is bad)} \end{cases} \end{cases}$$

The associated incremental individual utility in period 1 is:

$$\delta(t) = \begin{cases} \delta(0) < 0 \\ \delta(g) \ \geq 0 \\ \delta(b) \ = 0 \end{cases}$$

Therapy has a beneficial effect in the current period, whether or not it promotes resistance in the future; we assume that this effect is to maintain utility during the current period; not treating allows immediate degradation of a patient's condition. As discussed already, we focus on a situation where the difference between a good and a bad therapy does not lie so much in its current utility effect as in its future impact through the mechanism of viral resistance. If Therapy 1 proves effective (g), that beneficial effect will be felt mostly during the next period in the form of a better response to medication. As far as the current effect is concerned, it is at least as good when the current therapy is good as when it is bad; it may even be better. Hence our assumption that $\delta(g) \geq \delta(b)$.

Let $a(t)$ be the proportion of patients given Therapy 1 in period 1; since the treatment is chosen before the efficiency of Therapy 1 is revealed, $1 \geq a(g) = a(b) = 1 - a(0) \geq 0$. The total population of HIV infected patients is normalized to 1. The *ex ante* subjective probability that Therapy 1 is successful (g) is γ.

Clinicians have a wider set of decision possibilities to choose from at the beginning of period 2. They can abstain from any treatment; they can prescribe Therapy 1; or they can prescribe Therapy 2. Furthermore, as discussed in more details below, the effect of each treatment in period 2 depends on the treatment history of the patient (whether the patient has been administered Therapy 1 or has not been treated), and on the effectiveness of both treatments, that is on whether Treatment 1 is of type b or g (which is known at the beginning of period 2) and whether Therapy 2 is B or G (still unknown). Let T be the treatment undergone by a patient in period 2:

$$T = \begin{cases} 0 \text{ (no treatment in period 2)} \\ b \text{ (treatment with Therapy 1, if it has proven bad)} \\ g \text{ (treatment with Therapy 1, if it has proven good)} \\ T' = \begin{cases} = B \text{ (treatment with Therapy 2, if it proves bad)} \\ = G \text{ (treatment with Therapy 2, if it proves good)} \end{cases} \end{cases}$$

The proportion of patients who have been submitted to treatment t in period 1 and undergo treatment T in period 2 is $A(t, T)$; if patients from one single category are allowed to get different alternative treatments, then $A(t, T)$ is a random valued function. At the beginning of period 2, t is a known variable that can take three values, 0, b, or g; however, since the efficiency of Therapy 1 has been revealed, t actually can take only two values: 0 or g if Therapy 1 has proven good; 0 or b if Therapy 1 has proven bad. On the contrary, since the second-period treatment is chosen before the efficiency of Therapy 2 is revealed, the proportions A are such that $A(t, T') = A(t, B) = A(t, G)$ where T' is a random variable. Finally shares sum to unity. To sum up,

$$1 \geq A(t, B) = A(t, G) = A(t, T') = 1 - A(t, t) - A(t, 0) \geq 0; \; \forall t \tag{1}$$

$$1 \geq A(t, t) \geq 0; \; 1 \geq A(t, 0) \geq 0 \; \forall t \tag{2}$$

The associated incremental individual utility in period 2, discounted to period 1, $\Delta(t, T)$, is discussed in the next section. The problem faced by the clinician is a collective optimization problem, namely the maximization of expected incremental utility to the whole population of HIV-infected patients, minus total cost of treatments.[2] Because the second-period decision is conditional on the choice made in period 1, the solution may be obtained by backward induction. This means maximizing the net expected

[2] We abstract from considering the consequences of the HIV treatment decisions on other categories of patients.

incremental utility at the beginning of the second period for each possible state of the world associated with the initial choices of period 1. Once these various second-period programs have been solved, it becomes possible to optimize the first-period optimal *ex-ante* choice.

3. Second-period optimization

3.1 Utility effects of alternative period 2 treatments

At the beginning of period 2, the decision maker is fully informed on Therapy 1 and must make a decision on future treatment given the effectiveness of Therapy 1 and *a priori* beliefs on the effectiveness of Therapy 2, given the treatment administered to various patients in period 1. Consider first the case where Therapy 1 is ineffective: possible period 1 treatments are $t = 0$ or $t = b$. The incremental utility gains associated with the second-period choice of treatment, discounted to period 1, satisfy the following properties:

1. $\Delta(b, b) = \Delta(b, 0) \leq 0$: in case Therapy 1 has been applied $(t = b)$, the incremental gains associated with applying the same treatment 1 $(T = b)$, or no treatment $(T = 0)$, in the second period are non positive.

2. $\Delta(0, 0) = \Delta(b, 0) \leq 0 < \Delta(0, b)$: applying no therapy in period 2 implies the same utility loss whether the patient received Therapy 1 in the first period or not. However, if no treatment has been applied in period 1 $(t = 0)$, applying Therapy 1 in period 2 $(T = b)$ is more effective than no treatment $(T = 0)$; in other words, although Therapy 1 (b) loses its effectiveness after one period, applying it over the current period provides a gain if it has not been applied before; the same holds for Therapy 2 (see next item).

3. $\Delta(0, 0) \leq 0 < \Delta(0, B)$: if Therapy 2 is not successful, it yields a temporary improvement over no treatment if no treatment has been applied before.

4. $\Delta(b, B) = \Delta(b, b) \leq 0 < \Delta(0, b) = \Delta(0, B)$: if Therapy 2 is not successful (B), it is as good as, and does not yield any improvement over, Therapy 1.

5. $\Delta(b, G) < \Delta(0, G)$: if Therapy 2 is successful, it is less effective if the patient has been treated with Therapy 1 in the past $(t = b)$ rather than not treated $(t = 0)$.

6. $\Delta(0, G) \geq \Delta(0, B)$; $\Delta(b, G) \geq \Delta(b, B)$: if Therapy 2 is successful, it is more effective than if it turns out ineffective, whether or not the patient has undergone Therapy 1. (see also Property 9, applying when Therapy 1 is effective)

   Consider now the case where Therapy 1 is effective (g). The incremental utility gains associated with the second-period choice of treatment satisfy the following properties:

7. $\Delta(g, G) > \Delta(0, G) > 0$, $\Delta(g, B) > \Delta(0, B) > 0$, $\Delta(g, g) > \Delta(0, g) > 0$: earlier initiation of a good Therapy 1 improves the utility gain from either Therapy 1 or Therapy 2 in the second period, whether Therapy 2 turns out to be efficient or not.

8. $\Delta(g, 0) = \Delta(0, 0) \leq 0$: No treatment in period 2 implies the same utility loss, whatever the previous treatment. In order to take advantage of a 'good' Therapy 1, one must apply some therapy in period 2.

9. $\Delta(g, G) > \Delta(g, B) = \Delta(g, g) \geq \Delta(0, g) = \Delta(0, B) > 0$; $\Delta(0, G) > \Delta(0, g) = \Delta(0, B) > 0$: if Therapy 2 is not successful, it may be as good as, but does not yield any improvement over, Therapy 1; however if Therapy 2 is successful it yields a higher utility improvement than Therapy 1.

10. $\Delta(g, g) < \Delta(0, G)$: Therapy 2, if successful, is sufficiently superior to a successful Therapy 1 to dominate it even if the patient has not had the benefit of early treatment with Therapy 1.

3.2 The resistance effect

Several of the above properties result from the effect of period 1 treatment on viral resistance, an effect which has the nature of an opportunity cost. When Therapy 1 proves 'bad' the resistance to future medication is increased if Therapy 1 has been administered. As already mentioned in the introduction, this resistance effect is an empirical fact that has been demonstrated in the medical literature. In our model, it is measured by the difference between the utility change from Therapy 2 on 'naive' patients and the utility change from the same therapy on previously treated patients. Depending on whether Therapy 2 turns out 'good' or 'bad', this gives two possibilities:

$$R_G(0,b) \equiv \Delta(0,G) - \Delta(b,G) \tag{3a}$$

$$R_B(0,b) \equiv \Delta(0,B) - \Delta(b,B) \tag{3b}$$

Similarly, when Therapy 1 turns out 'good', the patient benefits from early administration of the therapy; resistance, and regret, will build up if Therapy 1 is not administered:

$$R_G(g,0) \equiv \Delta(g,G) - \Delta(0,G) \tag{4a}$$

$$R_B(g,0) \equiv \Delta(g,B) - \Delta(0,B) \tag{4b}$$

3.3 Therapy 1 has proven ineffective (b)

Conditional on Therapy 1 having turned out ineffective, and given period 1 treatment decision $a(t)$ $(t = 0$ or $t = b)$, the expected *ex-ante* net incremental utility of period 2 therapeutic choice depends on the treatment decisions applied at the beginning of the period, namely apply Therapy 1, whose realized value is b, apply Therapy 2, whose value B or G is still unknown, or not apply any therapy. This treatment decisions must be made for the group of previously treated patients, whose proportion is $a(b)$, and for the group of untreated patients, whose proportion is $a(0)$. Thus, conditional on Therapy 1 having proven ineffective, the *ex ante* net-of-treatment costs expected incremental utility is

$$
\begin{aligned}
V_b &\equiv E\left\{a(b)\,\Delta(b,T) + a(0)\,\Delta(0,T) - C(Z(b))\right\} \\
&= a(b)\left[A(b,b)\,\Delta(b,b) + A(b,G)\,\Gamma\Delta(b,G) + A(b,B)(1-\Gamma)\,\Delta(b,B) + A(b,0)\,\Delta(b,0)\right] \\
&\quad + a(0)\left[A(0,b)\,\Delta(0,b) + A(0,G)\,\Gamma\Delta(0,G) + A(0,B)(1-\Gamma)\,\Delta(0,B) + A(0,0)\,\Delta(0,0)\right] \\
&\quad - \frac{1}{2}C(Z(b))
\end{aligned}
$$

where Z is the number of patients undergoing some form of therapy in period 2; precisely, among the proportion $a(b)$ of patients who have undergone Therapy 1, some may be given Therapy 1 again while others may be given Therapy 2, and similarly for 'naive' patients: $Z(b) = a(b)(A(b,b) + A(b,T')) + a(0)(A(0,b) + A(0,T'))$, where T' is a stochastic variable that may take the values B or G. Since the second period choice of treatment occurs before the efficiency of Therapy 2 is known, the proportion of patients receiving Therapy 2 will be the same whether the therapy turns out to be effective or not: $A(b,T') = A(b,G) = A(b,B)$ and $A(0,T') = A(0,G) = A(0,B)$. Substituting, and using the fact that $a(0) = (1 - a(b))$,

$$
\begin{aligned}
V_b &= a(b)\left[A(b,b)\,\Delta(b,b) + A(b,T')(\Gamma\Delta(b,G) + (1-\Gamma)\,\Delta(b,B)) + A(b,0)\,\Delta(b,0)\right] \\
&\quad + (1-a(b))\left[A(0,b)\,\Delta(0,b) + A(0,T')(\Gamma\Delta(0,G) + (1-\Gamma)\,\Delta(0,B)) + A(0,0)\,\Delta(0,0)\right] \\
&\quad - \frac{1}{2}C\left[a(b)(A(b,b) + A(b,T')) + (1-a(b))(A(0,b) + A(0,T'))\right]^2
\end{aligned}
$$

This objective function must be maximized by choice of the proportions $A(t,T)$ of patients undergoing various treatments in period 2: $A(b,b)$ for patients who are kept on Therapy 1; $A(b,0)$ for patients who are taken out of Therapy 1; $A(0,0)$ for patients who continue receiving no therapy; $A(b,T')$ for patients who have undergone Therapy 1 and will receive Therapy 2, whether it proves good $(T' = G)$ or bad $(T' = B)$; $A(0,T')$ for patients who have never received any therapy and will receive Therapy 2.

Since, by Property 1, there is no gain from using b in period 2 if it has been used before, $A(b,b) = 0$; since, by Properties 4 and 6 $\Delta(0,T') \geq \Delta(0,b)$ whether $T' = G$ or B while treatments g or T' have the same cost, $A(0,b) = 0$. On the other side, not providing any therapy saves costs, so that the proportion of patients who do not undergo therapy cannot be set to zero on the ground that this does not provide utility to them. However, shares sum up to one so that $A(b,0) = 1 - A(b,T') - A(b,b)$; and $A(0,0) = 1 - A(0,T') - A(0,b)$. Making the substitutions, the net expected incremental period 2 utility, conditional on Therapy 1 having revealed itself ineffective (b), is

$$\begin{aligned} V_b &= a(b)\left[A(b,T')\left(\Gamma\Delta(b,G) + (1-\Gamma)\Delta(b,B) - \Delta(b,0)\right) + \Delta(b,0)\right] \\ &\quad + (1-a(b))\left[A(0,T')\left(\Gamma\Delta(0,G) + (1-\Gamma)\Delta(0,B) - \Delta(0,0)\right) + \Delta(0,0)\right] \\ &\quad - \frac{1}{2}C\left[a(b)\left(A(b,T')\right) + (1-a(b))\left(A(0,T')\right)\right]^2 \end{aligned}$$

Two shares remain to be selected. Defining $x \equiv A(b,T')$, $y \equiv A(0,T')$, treating V_b as function of x and y, and noting that $\Delta(b,0) - \Delta(0,0) = 0$ by Property 2, the problem is now:

$$\max_{x,y} aKx + (1-a)Jy + \Delta(0,0) - \frac{1}{2}C\left[ax + (1-a)y\right]^2 \tag{5}$$

where $K = \Gamma\Delta(b,G) + (1-\Gamma)\Delta(b,B) - \Delta(b,0)$; $J = \Gamma\Delta(0,G) + (1-\Gamma)\Delta(0,B) - \Delta(0,0)$; and the argument of a has been dropped. The first two terms represent expected utility payoffs while the third and fourth terms are certain utility change and cost.

Parameter K represents the expected utility gain from using Therapy 2, relative to using no therapy (i.e. relative to a utility loss of $\Delta(b,0)$), on patients that have been submitted to Therapy 1. Parameter J represents the expected utility from using Therapy 2, relative to using no therapy (i.e. relative to a utility loss of $\Delta(0,0)$), on patients that have not been submitted to any therapy before. It is noticeable that K represents a riskier option than J:[3] in case of success, the payoff is higher in the first option than in

[3] We ignore the non stochastic components $\Delta(b,0)$ and $\Delta(0,0)$ in the comparison.

the second one ($\Delta(b,G) > \Delta(0,G)$;) while, in case of a bad outcome, the first option implies a utility loss for previously treated patients while the second one implies a low, but non negative, utility gain for naive patients ($\Delta(b,B) \leq 0 < \Delta(0,B)$ by Property 4).

It is shown in the Appendix that a solution where both x and y would be interior requires:

$$\Gamma\Delta(b,G) + (1-\Gamma)\Delta(b,B) - \Delta(b,0) = \Gamma\Delta(0,G) + (1-\Gamma)\Delta(0,B) - \Delta(0,0). \tag{6}$$

In order both x and y to be interior, i.e. strictly between zero and one, the success probability, the therapy payoffs, and the utility losses in case of no-therapy, would have to be exactly such that it is indifferent whether a patient undergoes Therapy 1 or no therapy in period 1, given that Therapy 1 is bad. This would happen only by chance, considering that the decision maker cannot influence this condition by the earlier choice of a. It follows that at least one of x or y is a corner solution, i.e. has a value of 0 or 1.

There are two alternative *ex post* therapeutic situations:[4] $K < J$ i.e. the expected utility gain from Therapy 2 is higher for patients who have not undergone Therapy 1 in period 1 if the latter turns out 'bad', or the opposite. We relegate the analysis of the second alternative to the Appendix and focus here on the former because it corresponds to the debate about HAART therapies. Obviously, it can generate regret: if someone has received Therapy 1 and the latter turns out 'bad', then the decision-maker wished she had not prescribed Therapy 1. In that sense, J is a better *ex post* therapeutic option when Therapy 1 is 'bad', and must be preferred to K. If one of the two shares is set at one, then, necessarily, $y = 1$, thereby selecting all patients who have not undergone Therapy 1 as candidates for Therapy 2. In that instance the solution is found by setting $y = 1$ and then optimally choosing x within the interval $[0,1]$. This is optimal if $\frac{\partial V(x,1)}{\partial y} \geq 0$.[5]

[4] These therapeutic situations are called *ex post* because they are conditional on Therapy 1 having proven 'bad'.

[5] Since $\frac{\partial^2 J}{\partial x \partial y} = -c^2 a(1-a)$, the condition $\frac{\partial J(1,y)}{\partial x} > 0$ is met for all y if it is true at $y = 1$.

Alternatively, if one of the shares is set at zero, it must be x, with y chosen within the interval $[0,1]$. This is optimal if $\frac{\partial V(0,y)}{\partial y} < 0$.

As represented in Figure 1, the above argument shows that the solution to problem (5) lies on the perimeter of a square whose unit sides represent the respective values of x, the share of previously treated patients who receive Therapy 2, and y, the share of 'naive' patients who receive Therapy 2.

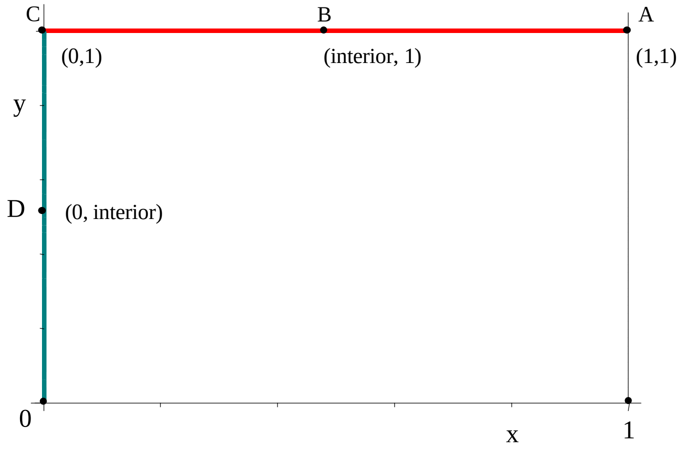

Figure 1: x proportion of previously treated patients undergoing Therapy 2; y proportion of 'naive' patients undergoing Therapy 2

In the case $K < J$, the above discussion also implies that $y \geq x$ since J is a better *ex post* option; consequently the solution lies on the east or north segments of the square, on the bold lines. Various possibilities are represented by points A $(x^* = 1,\ y^* = 1)$, B $(x^* \text{ interior},\ y^* = 1)$, C $(x^* = 0,\ y^* = 1)$, and D $(x^* = 0,\ y^* \text{ interior})$.

A visual inspection of (5) indicates that the solution depends on the cost parameter C, therapeutic characteristics represented by J and K, and the proportion a of patients administered Therapy 1 in the first period. Treating a as a parameter at this stage, the parameter constraints derived from the conditions on $\frac{\partial V(x,y)}{\partial x}$ and $\frac{\partial V(x,y)}{\partial y}$ that arise in each case, and from the conditions that both x and y belong to the interval $[0,1]$, are gathered in Table 1 and presented in Figure 2. The combined values of a, C, J, and K

that give rise to the solutions $x = A(b, T')$ and $y = A(0, T')$ represented by points A, B, C, and D in Figure 1 are the same that define the loci A_b, B_b, C_b, and D_b in Figure 2.[6]

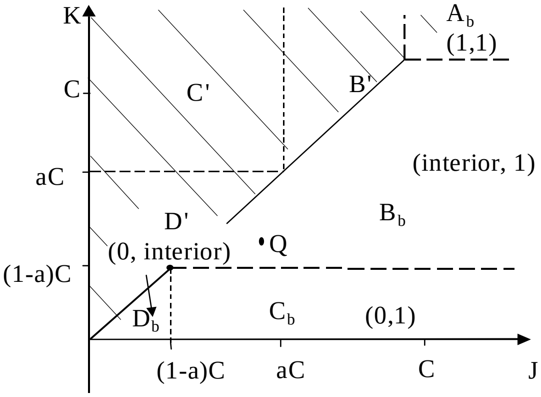

Figure 2: *ex post* contigent shares $\left(A\left(b, T'\right),\ A\left(0, T'\right)\right.$ according to therapy utility, cost, and period 1 share

For each set of conditions, Table 1 also gives the optimal period 2 shares (x^*, y^*), the corresponding locus in Figure 2, and the optimized value function $V^*\left(a|b\right)$ defined as[7] $V^*\left(a|b\right) \equiv V\left(x^*, y^*\right) = aKx^* + \left(1-a\right)Jy^* + \Delta - \frac{1}{2}C\left[ax^* + \left(1-a\right)y^*\right]^2$, with $\Delta \equiv \Delta\left(0, 0\right)$.

Table 1: Optimal period 2 shares and net utility when Therapy 1 proves ineffective and was less attractive than no therapy *ex ante* $(K < J)$

Conditions	Optimal period 2 shares (x^*, y^*)	locus	Optimized Value $V^*\left(a\|b\right)$
$C \leq K < J$	$(1,1)$	A_b	$a\left(K-J\right) + J + \Delta - \frac{1}{2}C$
$\left(1-a\right)C < K < C$	$\left(\frac{K-(1-a)C}{aC}, 1\right)$ with $0 < \frac{K-(1-a)C}{aC} < 1$	B_b	$\left(1-a\right)\left(J-K\right) + \Delta + \frac{1}{2}\frac{K^2}{C}$
$K \leq \left(1-a\right)C < J$	$(0,1)$	C_b	$\left(1-a\right)J + \Delta - \frac{1}{2}C\left(1-a\right)^2$
$K < J \leq \left(1-a\right)C$	$\left(0, \frac{J}{(1-a)C}\right)$ with $0 < \frac{J}{(1-a)C} < 1$	D_b	$\frac{1}{2}\frac{J^2}{C} + \Delta$

Locus A_b is particular in that its boundaries are not affected by any condition under the control of the decision maker: if both K and J are above C, the pair (J, K) rep-

[6] The dashed part of the figure represents the case where $J < K$. See Table A in the Appendix.

[7] The optimized value functions differ by the optimal solutions (x^*, y^*).

resented by any point in A_b remains in that locus whatever a; in other words, for any share of patients undergoing Therapy 1 in period 1, the solution in period 2 is the same: all patients undergo Therapy 2. This is not true of other loci such as B_b, C_b and D_b. By choosing a, the decision maker affects the position of the dashed $(1-a)\,C$ lines in Figure 2, which determines the boundaries of these loci. For example, if a is reduced, point Q, which belongs to locus B, will become an element of locus C, then of locus D, as the horizontal $(1-a)\,C$ line moves up and the vertical $(1-a)\,C$ line moves to the right.

We are now ready to define $V^*\,(a|b)$ according to parameter conditions that are independent of the variable a. Defining $\frac{K}{C}$ as the *marginal efficiency cost ratio of* K and $\frac{J}{C}$ as the *marginal efficiency cost ratio of* J, three different cases arise:

- Case b1: $1 \leq \frac{K}{C} < \frac{J}{C}$. The *marginal efficiency cost ratios* of both K (no therapy in period 1) and J (therapy in period 1) are above unity. The period 2 solution involves administering Therapy 2 to all patients irrespective of a; however, the value function depends on a: it is linearly decreasing:[8]

$$V^*\,(a|\,b) = a\,(K-J) + J + \Delta - \frac{1}{2}C$$

- Case b2: $\frac{K}{C} < 1 \leq \frac{J}{C}$. The *marginal efficiency cost ratio* of J (no therapy in period 1) is above unity; The *marginal efficiency cost ratios* of K (therapy in period 1) is below unity. Any (J, K) pair may be in locus C (low a; $x^* = 0$, $y^* = 1$) or in B (high a; $x^* = \frac{K-(1-a)C}{aC}$, $y = 1$). The critical frontier is when a is such that $K = (1-a)\,C$ (see Figure 1). In this intermediate case, the choice of a may affect x^*, but not y^* because the *efficiency cost ratio* of therapeutic option J triggered

[8] The difference between $K < J$ and $K > J$ arises here: in the first instance J^* is rising in a; in the second instance it is decreasing in a, with implications on the desirability to set $a = 1$ or $a = 0$ in period 1.

by y is very high. The optimized value function, as a function of a is:

$$V^*\left(a|\,b\right)=\begin{cases}\left(1-a\right)J-\frac{1}{2}C\left(1-a\right)^2+\Delta,\ 0\leq a\leq 1-\frac{K}{C}\\ \left(1-a\right)\left(J-K\right)+\frac{1}{2}\frac{K^2}{C}+\Delta,\ 1-\frac{K}{C}<a\leq 1\end{cases}$$

- Case b3: $\frac{K}{C}<\frac{J}{C}<1$. The *marginal efficiency cost ratios* of both K (no therapy in period 1) and J (therapy in period 1) are below unity. The period 2 solution involves three possibilities depending on the choice of a. As is clear from Figure 1, the critical frontier between loci D (low a) and C (intermediate a) satisfies $J=\left(1-a\right)C$, and the critical frontier between loci C and B (high a) satisfies $K=\left(1-a\right)C$. The optimized value function, as a function of a is:

$$V^*\left(a|\,b\right)=\begin{cases}\frac{1}{2}\frac{J^2}{C}+\Delta,\ 0\leq a\leq 1-\frac{J}{C}\\ \left(1-a\right)J-\frac{1}{2}C\left(1-a\right)^2+\Delta,\ 1-\frac{J}{C}<a\leq 1-\frac{K}{C}\\ \left(1-a\right)\left(J-K\right)+\frac{1}{2}\frac{K^2}{C}+\Delta,\ 1-\frac{K}{C}<a\leq 1\end{cases}$$

The value function $V^*\left(a|\,b\right)$ is represented in Figure 3 as a function of a for the three alternative cases just described.[9] Although it takes different forms according to relative utility impacts and cost of alternative treatments, the value function is in all instances concave and decreasing in the proportion a of patients submitted to Therapy 1 over the interval $[0,1]$. This property reflects the assumption $K<J$ that the expected payoff from Therapy 2 is higher when the therapy is applied to 'naive' patients. If Therapy 1 was certain to turn out 'bad', the proper period 1 decision would be to set $a=0$. However, as shown next, the second-period expected value function is different if Therapy 1 turns out 'good'.

[9] The parameters are $C=10$; $\Delta=0$. In case b1, $J=12$; $K=11.5$; in case b2, $J=11$; $K=7$; in case b3, $J=9$; $K=6$.

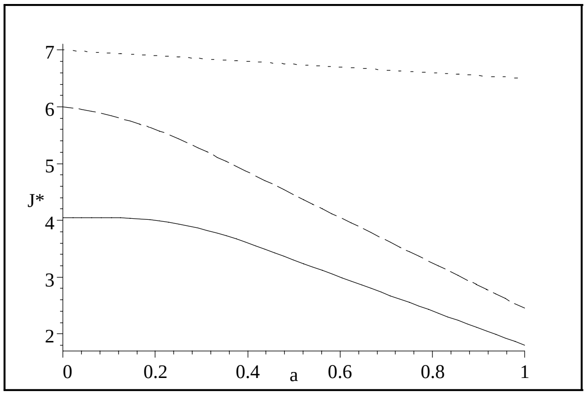

Figure 3: Optimized period 2 value function when Therapy 1 has proven bad and $K < J$(low *marginal efficiency cost ratio*: solid; medium ratio: dash; high ratio: dots)

3.4 Therapy 1 has proven effective (g)

In the alternative case, where Therapy 1 has proven effective, the *ex ante* net-of-treatment costs expected incremental utility for period 2 is

$$V_g \equiv E\left\{a(g)\,\Delta(g,T) + a(0)\,\Delta(0,T) - c(Z(g))\right\}$$

Since $A(g,T') = A(g,G) = A(g,B)$, $A(0,T') = A(0,G) = A(0,B)$, and $a(0) = (1 - a(g))$ we have

$$\begin{aligned} V_g &= a(g)\left[A(g,g)\,\Delta(g,g) + A(g,T')\left(\Gamma\Delta(g,G) + (1-\Gamma)\,\Delta(g,B)\right) + A(g,0)\,\Delta(g,0)\right] \\ &\quad + (1-a(g))\left[A(0,g)\,\Delta(0,g) + A(0,T')\left(\Gamma\Delta(0,G) + (1-\Gamma)\,\Delta(0,B)\right) + A(0,0)\,\Delta(0,0)\right] \\ &\quad - \frac{1}{2}C\left[a(g)\left(A(g,g) + A(g,T')\right) + (1-a(g))\left(A(0,g) + A(0,T')\right)\right]^2 \end{aligned}$$

where the number of patients undergoing some form of therapy in period 2 is $Z(g) = a(g)\left(A(g,g) + A(g,T')\right) + a(0)\left(A(0,g) + A(0,T')\right).$

Since, by Property 9, $\Delta(g,G) > \Delta(g,B) = \Delta(g,g)$ while treatments g or T' have the same cost, $A(g,g) = 0$; since, by Property 9 again, $\Delta(0,T') \geq \Delta(0,g)$ whether $T' = G$ or B, $A(0,g) = 0$. Thus Therapy 1 is never used in period 2. From these restrictions and the fact that shares sum up to one, it follows that $A(g,0) = 1 - A(g,T')$; and

$A(0,0) = 1 - A(0,T')$. Consequently,

$$\begin{aligned} V_g &= a(g)\left[A(g,T')\left(\Gamma\Delta(g,G) + (1-\Gamma)\Delta(g,B) - \Delta(g,0)\right) + \Delta(g,0)\right] \\ &\quad + (1-a(g))\left[A(0,T')\left(\Gamma\Delta(0,G) + (1-\Gamma)\Delta(0,B) - \Delta(0,0)\right) + \Delta(0,0)\right] \\ &\quad - \frac{1}{2}C\left[a(g)\left(A(g,T')\right) + (1-a(g))\left(A(0,T')\right)\right]^2 \end{aligned}$$

Two shares remain to be selected; defining $v \equiv A(g,T')$, $w \equiv A(0,\ T')$,[10] , and $V(v,w) \equiv V_g$; noting that $\Delta(g,0) - \Delta(0,0) = 0$ by Property 8, the problem is now:

$$\max_{v,w} aMv + (1-a)Lw + \Delta(0,0) - \frac{1}{2}C\left[av + (1-a)w\right]^2 \tag{7}$$

where $M = \Gamma\Delta(g,G) + (1-\Gamma)\Delta(g,B) - \Delta(g,0)$ and $L = \Gamma\Delta(0,G) + (1-\Gamma)\Delta(0,B) - \Delta(0,0)$. In complete similarity with the case of ineffective Therapy 1, the first two terms respectively represent the expected utility payoffs of applying Therapy 2 to previously treated patients, or to 'naive' patients, while the third and fourth terms are certain utility changes and costs.

It is now clear that the period 2 problem that arises when Therapy 1 proves efficient has exactly the same structure as when Therapy 1 proves inefficient. The two problems differ only in the decision variables (v and w versus x and y) and in the parameters (M and L versus K and J).

Consequently, a similar structure, involving three different cases, emerges for each of the main therapeutic situations $(M < L$ and $L < M)$, according to the *marginal efficiency cost ratios.* When Therapy 1 is efficient, the marginal efficiency cost ratios are defined as $\frac{M}{C}$ for patients administered Therapy 1 in period 1 and $\frac{L}{C}$ for naive patients. The case $L < M$ corresponds closely to HAART therapies in that, whatever happens

[10] w is not the same variable as y, which applies when Therapy 1 is 'bad'. As is obvious from the way we are constructing the complete solution, the choice of shares $A(.,.)$ in period 2 is conditional on whether Therapy 1 has proven 'good' or 'bad'. This appears clearly in the notation $A(g,.)$ or $A(b,.)$ but not in $A(0,.)$. A proper notation could be $A(0|g,.)$ rather than $A(0,.)$. Since we substitute the optimized value of the A's in the continuation of the treatment we do not make the distinction for notational simplicity's sake, except in Table 2, where the optimal shares are presented for completeness.

to Therapy 2, that therapy will be more efficient on patients who have benefitted from Therapy 1 previously. It will be treated in detail. As when Therapy 1 turns out 'bad', there are three possible cases:

- Case g1: $1 \leq \frac{L}{C} < \frac{M}{C}$. The *marginal efficiency cost ratios* of both L (no therapy in period 1) and M (therapy in period 1) are above unity. The period 2 solution involves administering Therapy 2 to all patients irrespective of a; despite the independence of the solution on a, the optimum value function is linearly increasing:

$$V^*\left(a|\, g\right) = a\left(M - L\right) + L + \Delta - \frac{1}{2}C$$

- Case g2: $\frac{L}{C} < 1 \leq \frac{M}{C}$. The *marginal efficiency cost ratio* of M (therapy in period 1) is above unity; The *marginal efficiency cost ratios* of L (no therapy in period 1) is below unity. In this intermediate case, the choice of a may affect w^*, but not v^* because the *efficiency cost ratio* for patients having undergone Therapy 1 is very high. The optimized value function, as a function of a is rising and linear for low a's, then rising at a decreasing rate:

$$V^*\left(a|\, g\right) = \begin{cases} a\left(M - L\right) + \Delta + \frac{1}{2}\frac{L^2}{C},\ 0 \leq a \leq \frac{L}{C} \\ aM + \Delta - \frac{1}{2}Ca^2,\ \frac{L}{C} < a \leq 1 \end{cases}$$

- Case g3: $\frac{L}{C} < \frac{M}{C} < 1$. The *marginal efficiency cost ratios* of both L (no therapy in period 1) and M (therapy in period 1) are below unity. The period 2 solution involves three possibilities depending on the choice of a. The optimized value function is rising linearly at low values of a, then rising at a decreasing rate, and finally constant:

$$V^*\left(a|\, g\right) = \begin{cases} a\left(M - L\right) + \Delta + \frac{1}{2}\frac{L^2}{C},\ 0 \leq a \leq \frac{L}{C} \\ aM + \Delta - \frac{1}{2}Ca^2,\ \frac{L}{C} < a \leq \frac{M}{C} \\ \frac{1}{2}\frac{M^2}{C} + \Delta,\ \frac{M}{C} < a \leq 1 \end{cases}$$

The above analysis confirms the initial intuition: if Therapy 1 proves effective, it is preferable, from an ex post point of view, to deal with patients who have received it. This does not mean that the right ex ante decision is to administer Therapy 1 to all patients: first the therapy may prove ineffective; second it involves costs in the first period.

4. First period optimization

4.1 *Ex ante* period 2 payoff

Period 1 treatment decision is based on the summation of period 1 payoffs and period 2 payoffs. The period 2 expected welfare from the treatment decision of period 1 is a function of the proportion a of patients given Therapy 1 and depends on γ, the *ex ante* probability of success of Therapy 1:

$$U\left(a\right)=\gamma V^{*}\left(a|g\right)+\left(1-\gamma\right)V^{*}\left(a|b\right)$$

where $V^{*}\left(a|b\right)$ depends on parameters J, K, and C while $V^{*}\left(a|g\right)$ depends on L, M, and C.

From the definitions of J, K, L, and M, and from the properties of the function $\Delta\left(.,.\right)$, $J=L$ and $M>K$; for the HAART case we have also argued that $K<J$ and $L<M$. As a result the *ex ante* therapeutic configuration pertaining to HAART is:

$$K<J=L<\ M$$

This set of inequalities summarizes the *ex ante* medical dilemma: is it preferable or not to administer Therapy 1 as early as period 1? From a strict therapeutic point of view, if Therapy 1 is administered in period 1 and later turns out 'bad' then the decision maker had rather abstained from administering it because of the emergence of viral strains resistant to treatment. However, if Therapy 1 turns out 'good' and has not been administered in period 1, the medical benefits from Therapy 2 are diminished because Therapy 1 would have slowed down the weakening of the immune system.

Furthermore, economic considerations complicate the decision. The lower the *marginal efficiency cost ratio*, the less desirable the therapy from an economic point of view. Since total cost is $\frac{1}{2}CZ^2$, where Z is the total proportion of patients undergoing some type of therapy, C represents the marginal cost of therapy when Z is at its maximum value of one. When not all patients undergo some form of therapy, Z is lower than unity, and the marginal cost is lower than C. Thus a *marginal efficiency cost ratio* in excess of unity means that the additional benefits from the therapy under consideration exceed its additional costs when $Z = 1$, but not necessarily at lower levels of Z. Precisely, $Z = \frac{1}{2}C\left[av + (1-a)\,w\right]^2$ if Therapy 1 turns out 'good' $(Z = \frac{1}{2}C\left[ax + (1-a)\,y\right]^2$ if Therapy 1 turns out 'bad') so that the marginal effect of changing a is smaller than C; it depends crucially on the *ex post* decisions studied earlier, and depends on the success probabilities as well.

We will focus on the difficult but interesting and realistic situation where the therapeutic dilemma is combined with a non trivial economic choice:

- a very promising therapeutic opportunity for patients submitted early to Therapy 1 in case both therapies turn out 'good': $1 < \frac{M}{C}$.
- in case one therapy turns out 'bad' or both do, *marginal efficiency costs ratios* below unity $(\frac{K}{C} < \frac{J}{C} \leq 1)$, but reasonably close to unity $(1 - \frac{J}{C} \leq 1 - \frac{K}{C} < \frac{K}{C} < \frac{J}{C})$ so that the corresponding therapy options remain attractive for some values of Z.[11]

This corresponds to the *ex post* cases $g2$ and $b3$ described in the previous sections: administering Therapy 2 to patients who have undergone Therapy 1 is automatic in case

[11] It turns out that it does not matter whether $1 - \frac{K}{C} < \frac{K}{C}$ or not provided $1 - \frac{J}{C} \leq 1 - \frac{K}{C} < \frac{J}{C} \leq 1 < \frac{M}{C}$. Treating other cases is just a matter of adapting the methodology. We leave it to interested readers to do so.

Therapy 1 has turned out 'good' (case $g2$), but not in other situations ($b3$), so that the *ex ante* decision on the administration of Therapy 1 remains complex in that sense.

Table 2 gives the optimal period 2 contingent treatment decision rules $A(.,.)$ in Column 2, and the expected period 2 payoff in Column 3.

Table 2. *Ex ante* period 2 contingent rules (x^*, y^*, w^*, v^*) and payoff $U(a)$

$x = A(b, T')\,;\ y = A(0|b, 0)\,;\ w = A(0|g, 0)\,;\ v = A(g, T')$

% given Therapy 1	(x^*, y^*, w^*, v^*)	$U(a) = (1-\gamma) V^*(a\vert b) + \gamma V^*(a\vert g)$
$0 \leq a \leq 1 - \frac{J}{C}$	$\left(0, \frac{\frac{J}{C}}{1-a}, \frac{\frac{J}{C}-a}{1-a}, 1\right)$	$\gamma a(M-J) + \Delta + \frac{1}{2}\frac{J^2}{C}$
$1 - \frac{J}{C} < a \leq 1 - \frac{K}{C}$	$\left(0, 1, \frac{\frac{J}{C}-a}{1-a}, 1\right)$	$(1-\gamma)\left((1-a)J - \frac{1}{2}C(1-a)^2\right) + \gamma\left(a(M-J) + \frac{1}{2}\frac{J^2}{C}\right) + \Delta$
$1 - \frac{K}{C} < a \leq \frac{J}{C}$	$\left(\frac{\frac{K}{C}-1+a}{a}, 1, \frac{\frac{J}{C}-a}{1-a}, 1\right)$	$(1-\gamma)\left((1-a)(J-K) + \frac{1}{2}\frac{K^2}{C}\right) + \gamma\left(a(M-J) + \frac{1}{2}\frac{J^2}{C}\right) + \Delta$
$\frac{J}{C} < a \leq 1$	$\left(\frac{\frac{K}{C}-1+a}{a}, 1, 0, 1\right)$	$(1-\gamma)\left((1-a)(J-K) + \frac{1}{2}\frac{K^2}{C}\right) + \gamma\left(aM - \frac{1}{2}Ca^2\right) + \Delta$

4.2 Total two-period net utility

The sum of period 1 and period 2 net expected payoffs is

$$W(a) = u(a) + U(a) \tag{8}$$

where $u(a)$ is the sum of utilities experienced in period 1 by the proportion a of patients administered Therapy 1 and the proportion $(1-a)$ that do not receive any therapy, net of the cost of treatment in period 1:

$$u(a) = a(\gamma\delta(g) + (1-\gamma)\delta(b)) + (1-a)\delta(0) - \frac{1}{2}ca^2 \tag{9}$$

The total payoff function is presented graphically in Figure 4.[12]

[12] Period 1 parameters are $c = .1$, $\gamma = .5$; period 2 parameters are $C = 10$; $\Delta = 0$; $K = 6$; $J = 9$; $M = 11$. They correspond to cases b3 and g2 of the period 2 analysis.

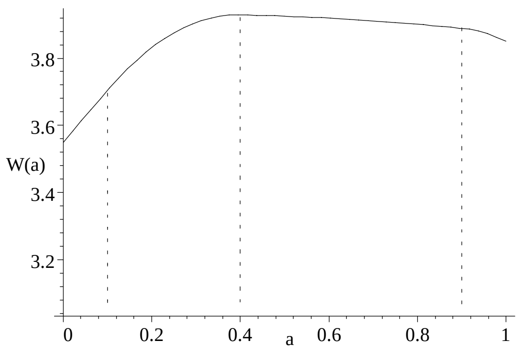

Figure 4: Intertemporal net expected utility $\left(1-\frac{J}{C}\leq 1-\frac{K}{C}<\frac{J}{C}\leq 1<\frac{M}{C}\right)$

Figure 4 illustrates an interior solution for a. While some of the curve's concavity is due to period 1 contribution to net utility, most of its qualitative properties arise from period 2 net expected utility. As indicated in Table 2, the latter is made up of four segments, defined over the intervals showed on the graph. These segments correspond to different combinations of the four possible period 2 optimal contingent treatments $(A^*(b,T'),\ A^*(0|b,0),\ A^*(0|g,0),\ A^*(g,T'))$. The slope of U is a continuous function of a, a smoothness due to the fact that some optimum period 2 shares are interior over some ranges.

4.3 The determination of first-period treatment

The first-period share a that maximizes $W(a)$ may be a corner solution or an interior one. Our analysis determines under what therapeutic and cost parameter combinations this occurs. A corner solution, $a=0$ or $a=1$, will not be sensitive to changes in parameters over some range. On the contrary, as explained above, the combination of a therapeutic dilemma with significant, but not prohibitive, economic costs, makes an interior solution more likely. From both the medical and the economic point of views, the first-period treatment decision is then most difficult and interesting.

When the solution is interior its properties can be analyzed from the first-order condition $\frac{\partial W}{\partial a}=0$. This condition may be satisfied on any of the four a intervals on which $W(a)$ is defined. The properties of the solution may differ accordingly. We show in the Appendix that the optimal proportion of patients administered Therapy 1 is, among the four possible values of a obtained by setting $\frac{\partial W}{\partial a} = 0$, the one which belongs to the corresponding interval of definition of $W(a)$ [13]

$$a = \begin{cases} \frac{n+\gamma(M-J)}{c} & \text{if compatible w.} \quad 0 \leq a \leq 1 - \frac{J}{C} \\ \frac{n+(1-\gamma)C+\gamma(M-J)-(1-\gamma)J}{c+(1-\gamma)C} & \text{if compatible w.} \quad 1 - \frac{J}{C} < a \leq 1 - \frac{K}{C} \\ \frac{n+K+\gamma(M-K)}{c} & \text{if compatible w.} \quad 1 - \frac{K}{C} < a \leq \frac{J}{C} \\ \frac{n+\gamma M-(1-\gamma)(J-K)}{c+\gamma C} & \text{if compatible w.} \quad \frac{J}{C} < a \leq 1 \end{cases} \tag{10}$$

where $n \equiv \gamma(\delta(g) - \delta(0)) + (1-\gamma)(\delta(b) - \delta(0))$ is the expected period 1 utility gain from applying Therapy 1 rather than no therapy; n measures the current effect, but not the resistance effect induced in the second period by the use of Therapy 1 in the first period.

For the parameter combination used in Figure 4, the solution is $a^* = 0.39$: this value of a, obtained from the second line of (10), also lies in the interval $\left[1 - \frac{J}{C},\ 1 - \frac{K}{C}\right]$. The rule adopted by the decision maker for future period 2 contingent choices can be traced back by referring to the period 2 results established above for the relevant a interval: for $1 - \frac{J}{C} < a \leq 1 - \frac{K}{C}$, Column 2 of Table 2 gives $A(b, T') = 0$ and $A(0|b, 0) = 1$ in case Therapy 1 turns out 'bad'; and $A(0|g, 0) = \frac{\frac{L}{C} - a}{1-a}$ and $A(g, T') = 1$ in case Therapy 2 turns out 'good'.

The four possible forms given for a in (10) indicate that the optimum proportion of patients submitted to Therapy 1 depends on n the direct utility gain from that therapy in period 1, on the utility gains J, K, and M that are made possible in period 2 by the use of Therapy 1 in period 1, and on the costs c and C of applying therapy in period 1 and some therapy in period 2. Since they are not observed *ex ante*, period 2 costs and utility gains are weighted according to the probability γ that Therapy 1 turns out 'good'. Some of these determinants of the optimal choice deserve comments.

[13] Since $W(a) = u(a) + U(a)$, the intervals of definition of $W(a)$ are inherited from those of $U(a)$, given in the first column of Table 2.

5. Discussion

5.1 Costs

Period 1 cost parameter c affects the decision to use Therapy 1 in the expected way. The second-period cost parameter has a more complex influence. In fact C does not enter lines 1 nor line 3 of expression (10). It would be mistaken to conclude that, over that interval, the number of patients given Therapy 2, hence the cost of that therapy, does not depend on the pool of individuals who have received Therapy 1. In fact one can see (Table 2, Column 2) that $A(0|b,0) = \frac{\frac{J}{C}}{1-a}$ and $A(0|g,0) = \frac{\frac{L}{C}-a}{1-a}$ for the a interval corresponding to line 1 in (10). This means that a not only affects period 2 shares, but also that the implied probability weighted cost for these patient categories is a function of a; however, the optimal contingent rule to be applied in period 2 is so designed that the contribution of cost to net payoffs in the second period, once the expected therapeutic effects are taken into account, is $\frac{1}{2}\frac{J^2}{C}$, which is independent of a (Column 3). Similarly, corresponding to line 3 of (10), two of the contingent shares given in Column 2 are interior $\left(A(b,0) = \frac{\frac{K}{C}-1+a}{a}\right.$ and $\left.A(0|g,0) = \frac{\frac{L}{C}-a}{1-a}\right)$, while the second period cost parameter C enters the value function as $(1-\gamma)\frac{1}{2}\frac{K^2}{C} + \gamma\frac{1}{2}\frac{J^2}{C}$, which is independent of a (Column 3).

In contrast, for the a intervals corresponding to lines 2 and 4 of (10), where at most one of the four contingent period 2 shares given in Column 2 of Table 2 is interior, the second period cost parameter affects the optimal choice of period 1 share. This points to a property of the optimum period 2 decision rule. If there is enough flexibility to choose an interior value for one of the period 2 shares in each possible realizations of Therapy 1 ('good' or 'bad'), then the decision maker will ensure that the expected net payoff in period 2 is not affected by the choice of a in period 1. In turn, the optimal value of a will then be disconnected from period 2 costs.

However it is not generally the case that two contingent shares are interior. In the configuration illustrated in Figure 4 (a^* given by line 2 of (10)), three of the optimal contingent shares are corner solutions while only one is interior: $A\left(b,T'\right)=0,\ A\left(0|b,0\right)=1,\ A\left(0|g,0\right)=\frac{\frac{L}{C}-a}{1-a}$, and $A\left(g,T'\right)=1$. If Therapy 1 turns out 'bad', no patient, whether he has undergone Therapy 1 or not, will be administered Therapy 2; the cost of Therapy 2 is then irrelevant. However, in case of a 'good' outcome, one of the contingent shares is interior while the other share is set at 1: the relationship between the decision made in period 1 and the net costs to be supported in period 2 is not severed in that case.

Corner solutions may be thought of as easy decisions in the sense that, at least over some parameter range, they are not sensitive to parameter variations. However the result just described implies that the easier and clear-cut the period 2 rule, the more forward looking and sophisticated the decision must be in period 1 because there will be little or no adjustment to its consequences.

5.2 Marginal efficiency cost ratios

While our discussion of costs has focused sofar on the absolute value of the cost parameters, the step by step description of the solution methodology has made it plain that the decision is structured according to *marginal efficiency cost ratios.* In fact the intervals of validity of the various forms of the first-period solution are defined in terms of *marginal efficiency cost ratios* only, and the same is true of the intervals defining the four period 2 contingent shares (Table 2, Column 2). The solution strategy can be interpreted as applying the following principle: choose a in such a way as to maximize future exposure to good outcomes and minimize exposure to bad outcomes.

To illustrate by way of the solution corresponding to Figure 4 again, where a^* is given by the second line in (10), consider the four optimal second-period contingent shares in that instance, as given on the second line (below headers) of Table 2: $\left(A^*\left(b,T'\right)=0; A^*\left(0|b,0\right)=1; A^*\left(0|g,0\right)=\frac{\frac{J}{C}-a}{1-a}; A^*\left(g,T'\right)=1\right)$. In case of a bad outcome b for Therapy 1 (first two contingent shares), the expected net-of-cost utility gain from Therapy 2 would be negative for previously treated patients: minimal exposure is achieved by not administering any therapy; on the contrary, the net expected gain of administering Therapy 2 to 'naive' patients is positive for any a : maximum exposure is achieved by administering the therapy to all 'naive' patients. In case of a good outcome g for Therapy 1 (second two contingent shares), the net expected gain of administering Therapy 2 to previously treated patients is positive for any a : maximum exposure is achieved by administering the therapy to all such patients; however, in case of a bad outcome for Therapy 2 the marginal utility gain may exceed marginal cost if the number of treated patients is low enough, but will fall short of marginal cost if too many patients get Therapy 2: maximum exposure to positive net benefits is achieved by setting the cutoff level of $A^{*}(0|g, 0)$ in such a way that society benefits from the policy; patients in excess of that cut-off level are not given access to the therapy, although they would privately benefit in the same way as patients in the same group who get the therapy.

5.3 Public goods and unequal treatment of equals

Any interior solution in our model is an instance of unequal treatment of equal individuals. This is a frequent implication of using a collective objective in decisions affecting individuals. Collective decision making, and health care decisions in particular, provide many instances of similar situations. As can be observed in the brief survey on standards and therapeutic criteria provided further below in Section 6., the use of additional health criteria may be subject to much controversy but has the advantage of shifting the issue, perhaps with some hypocrisy, from moral grounds back to scientific ones.

In our model, unequal treatment of equals results from the fact that private cost of therapy differs from the collective cost. As we have argued at the beginning of the paper, the assumption that total costs are not linear in the number of patients given therapy is probably realistic. However it is easy to adapt our model to other instances that create non linearities. A very interesting one is the externality involved when administering Therapy 1 to one patient potentially affects the efficiency of Therapy 2 on all patients. Thus, if the early administration of Therapy 1 to one patient not only affects his resistance, but may also affect the preponderance of more resistant viral strains that may be transmitted to other individuals, then the cost of Therapy 1 goes beyond the administration cost and affects other patients' utilities. This means that the number of patients administered a therapy affects the net benefits to all, in a way which is formally similar to what happens in our model, where the transmission mechanism is the direct total cost of therapy. Our model can be adapted to study that phenomenon specifically.

5.4 Relevant alternatives

The choice of a is affected by the current, direct, net utility gains from Therapy 1, but also by the net gains that may be expected over the next period. In applying the principle 'maximize exposure to good outcomes and minimize exposure to bad ones', the optimal period 2 contingent shares actually pick the relevant alternative treatment strategies and eliminate some others according to parameter values. For example, when $0 \leq a \leq 1-\frac{J}{C}$, the expected net period 2 payoff is $\gamma\frac{n+\gamma(M-J)}{c}(M-J)+\Delta+\frac{1}{2}\frac{J^2}{C}$ (Table 2, Column 3). Designating treatment strategies by their expected payoffs, the only relevant strategies are M and J (apply Therapy 2 to patients having undergone Therapy 1 and to naive patients, in case Therapy 1 turns out 'good'). Alternative strategy K (apply Therapy 2 to patients having undergone Therapy 1, in case Therapy 1 turns out 'bad') is simply not relevant over that interval. The same strategies M and J, not K, are relevant over the second interval, $\left(1-\frac{J}{C},\, 1-\frac{K}{C}\right)$, although they enter the value function in a different way, as their probabilities of being used differ.

However, for larger values of a (lines 3 and 4), strategy K becomes relevant. This happens as follows. As a increases, the number of naive patients $(1-a)$ becomes smaller. Since the totality of 'naive' patients receive Therapy 2 when Therapy 1 turns out 'bad' $(A(0|b,T')=1$ in lines 2, 3, and 4 of Column 2) while only part of those already treated receive the therapy $(A(b,T')=0$ in lines 1 and 2; $A(b,T')<1$ in lines 3 and 4), the total number of patients receiving Therapy 2 diminishes as a increases, so that the cost per patient diminishes (by the convexity of the cost function). Consequently strategy K becomes attractive, as indicated by the fact that $A(b,T')$ is interior over the interval corresponding to lines 3 and 4, instead of being set at zero, as in lines 1 and 2.

5.5 Resistance to therapy

As mentioned at the beginning of this paper, a major issue in HAART therapy is the development of resistance to treatments. This is illustrated in the forms taken by the period 2 payoff function of Table 2. This function is written in terms of the payoffs from treatment strategies J, K, and M. In each line, the expected payoff function include terms in the absolute levels of J, K, or M, but also terms measuring differences: $M-J$ and $J-K$. As will become clear when we substitute in the definitions of J, K, and M, these terms measure resistance effects. From the definitions of J, K, and M associated with (5) and (7), and from the definitions ($3b$) and ($4b$) of resistance effects from using or not using Therapy 1,

$$M-J = \Gamma R_G(g,0) + (1-\Gamma) R_B(g,0) \tag{11}$$

$$J-K = \Gamma R_G(0,b) + (1-\Gamma) R_B(0,b) \tag{12}$$

Thus, for low values of a, as in Line 1 of Table 2, the term $\gamma a (M-J)$ of Column 3 measures resistance. Expression (11) indicates that resistance is an issue in case Therapy 1 turns out 'good', and that potential resistance effects from not using Therapy 1 must be weighted according to the probabilities that Therapy 2 turns out 'good' or 'bad'.

In Line 3, terms in $J-K$ and $M-J$ enter the period 2 payoff function; this indicates that both resistance effects from using Therapy 1 when it should not be used (when Therapy 1 turns out 'bad'), and resistance effects from not using it when it should be used (when Therapy 1 turns out 'good'), are taken into account in the optimal decision, and properly weighted by the success probabilities of both therapies.

There is more to be observed about resistance effects. Consider again the interior solution illustrated in Figure 4, corresponding to the second line (below headers) of Table 2, and to the second line in (10). Other things equal, if the resistance induced by Therapy 1 in case the latter turns out 'bad' is higher, that is if K is lower and J unchanged, the domain of that solution $\left[1-\frac{J}{C}, 1-\frac{K}{C}\right]$ becomes larger, as illustrated in Figure 5 by the shift to the right of the vertical dashed line $1-\frac{K}{C}$. But, over that interval, a^* is insensitive to $J-K$ the resistance effect induced by submitting a patient to Therapy 1 if the latter turns out 'bad'. In other words, paradoxically, when the resistance effect is high, it becomes irrelevant to the choice of a over a larger range of possible a values.[14]

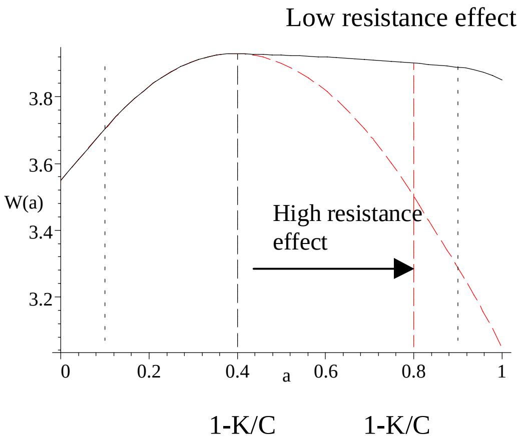

Figure 5: An increase in the Resistance effect (reduced K)

5.6 Real options

The decision to administer HAART is clearly a real-options problem. As in such problems, the decision maker here not only encounters costs but also looses some flexibility when she prescribes Therapy 1, and again looses some flexibility when she makes the period 2 decision. This is done in exchange not only for the benefits of Therapy 1, but also, in some parameter configurations, for the option to use Therapy 2 more efficiently if Therapy 1 turns out 'good', and for the option to abstain from Therapy 2 if Therapy 1 turns out 'bad'.

[14] The continuous expected value curve in Figure 5 is the same as in Figure 4; The dashed expected value curve has the same parameters as the continuous curve, except for K, which is set at .2 instead of .6, which implies a higher resistance effect.

As is frequently observed in such contexts (Dixit and Pindyck, 1994, especially chapters 1 and 2), allowing the period 2 decision to be taken when the relevant information has been obtained, rather than making an early decision based on expectations, allows the decision maker to retain some flexibility and exploit that flexibility later to increase the efficiency of the overall therapeutic sequence, or, in case of bad luck, to limit the cost of *ex post* mistakes. This is clearly a better strategy than maximizing the expected net payoff of choosing both a and A in period 1, as implied by standard expected net present value maximization.

Most real-options models differ from our model in that they adopt a continuous-time approach and focus on the critical value of one variable (sometimes two) at which the optimal decision changes from "wait" to "stop waiting". In our context, "stop waiting" could mean "initiate therapy" and the variable to watch in continuous time could be the probability of successful therapy. Considering only two periods allows us to introduce more sophistication in the decisions: the proportions administered a therapy are not constrained to be zero or one, and the decisions are conditional on patients' histories rather than simply the current state of a variable such as the probability of success.

6. Conclusion and limitations

The additional gains in life expectancy associated with HAART combinations are not yet precisely known. Studies that attempted to model the impact of antiretroviral treatments have provided estimations in the range of 6 to 24 years for the increase in an individual's life expectancy from HIV infection to death for patients treated in North America (Blower, Gershengorn & Grant, 2000 ;Wood et al., 2000a).

However, uncertainties remain about the long-term efficacy of HAART therapies. For example, recent evidence has shown the persistence of viral replication even in successfully treated patients (Zhang et al., 1999; Finzi et al., 1999). In such context it is not surprising that complete consensus has not been reached world-wide, among clinicians and health authorities, about the best standards of practice for HAART delivery. In particular, clinical guidelines still differ between countries, and sometimes inside each country, about the eligibility criteria for HAART initiation. Empirical studies draw conflicting conclusions on the matter: some of them strongly question the current aggressive early use of HAART (Tebas et al., 2001), whereas others advocate very early initiation (as soon as HIV-infected patients have less 500 CD4 cells/microL) on cost-effective grounds (Schackman, 2001). Interestingly, the former try to account for the resistance effect while the latter ignore it. Moreover, whatever the guidelines, surveys among HIV/AIDS prescribing physicians show a great variability of attitudes toward initiation of HAART treatment (Obadia et al., 1999 ; Reedjik et al., 1999 ; Kitahata, Van Rompaey & Shields, 2000 ; Landman et al., 2000).

In developing countries, where the vast majority of HIV infected people currently live, access to antiretroviral treatment was not considered a feasible technical and economic option until recently (Van Praag et al., 1997; Ainsworth and Teokul, 2000). Following the United Nations General Assembly Special Session on AIDS in 2001, a multi-lateral Global Fund to Fight AIDS, Tuberculosis and Malaria has been established at the beginning of 2002, and the goal of scaling up access to HAART in developing countries is increasingly shared by governments and international donor organisations. Between 1996 and 2000, expensive drug costs were the major barrier for diffusion of HAART in these countries. In the last three years, significant reductions in the prices of anti-retroviral and other HIV-related drugs have been brought about in developing countries with the greatest need for access to HAART. The debate about the rationality of promoting access to HAART in the developing world therefore tends to focus on other issues such as deficiencies in existing health care infrastructures (Sonnabend, 2000), the cost-effectiveness ratio of HAART in comparison with alternative use of scarce resources to improve public health (Creese, 2002; Marseille et al., 2002). Our simple stochastic dynamic model of a sequential therapeutic choice with intermediary revelation of information underlines the importance of expectations about effectiveness and efficiency over cost ratios in current and future therapies, as well as the importance of induced viral resistance. Scaling up access to HAART will require further reductions in costs of delivery that would have to be obtained not only through more affordable prices for all drugs entering in the various HAART regimens, but also by making available cheaper alternative techniques for biological monitoring of viral load and CD4 cell counts, and by "adapting" treatment and monitoring guidelines that have been initially developed in the North for wider use in resource-limited settings (Hammer et al., 2002).

Indeed, the fear that diffusion of HAART may spread viral resistance tends to become the most powerful argument in favour of limiting or delaying access to antiretroviral treatment. At the empirical level, there is evidence from the Brazilian programme of universal coverage for HAART and from pilot experiments in African countries such as Senegal and Uganda that viral resistance and non-adherence are not a greater problem in cohorts of patients treated in developing countries when compared to data from developed countries (Tanuri et al., 2002; Silveira et al., 2002; Weidle et al., 2002; Laurent et al., 2002). In any case, our model also shows that unilateral attitudes and arguments, such as the ones recommending to withhold or delay access to HAART in certain groups of patients or countries (Senak, 1997 ; Stewart, 1997 ; CDC, 1998) for fear of possible diffusion of drug-resistant HIV strains, express very questionable implicit trade-offs.

Moreover, a main conclusion of our model is that, when there is a significant risk of resistance due to therapeutic failure of initial existing treatments, differences in the estimation of this risk should not influence the optimal decision about the size of the HIV-infected population eligible for initiation of HAART. If this corresponds to the situation in developing countries (or in some marginalized groups of developed countries), priority should be given to pure efficiency over cost considerations in choosing criteria for initiating treatment (levels of CD4 cell counts, viral load and clinical stage of HIV infection). Paradoxically, it is in the case where expectations about resistance are rather optimistic (the phenomenon will be limited) that differences in estimations of this phenomenon may be a factor of variability in optimal treatment initiation. Because these conclusions are quite counter-intuitive, they may help clarify current inconsistencies between recommendations and practical behaviors of HIV/AIDS clinicians and public health experts on the one hand, and the expressed set of preferences and expectations of these same decision-makers, on the other hand (Gerbert et al., 2000).

A major limitation of our model is that we focus on the impact of the decision to initiate treatment on a population which is already HIV-infected. From a public health perspective, negative externalities associated with the diffusion of resistant HIV-strains as well as positive externalities related to effective treatment are important factors (Geoffard, Philipson, 1996). As we have mentioned earlier, our model can easily be adapted to take these externalities into account.

Finally, empirical validation is now needed to derive more detailed practical recommendations from our analysis. Future empirical attempts to model the potential demographic, epidemiological and economic impact of alternative scenarios for diffusion of HAART therapies in different HIV-infected populations could be improved by the use of a similar stochastic dynamic approach. Such an approach would also be useful in taking into account irreversibility and inertia effects associated with initial treatment choices.

REFERENCES

AINSWORTH, M. and TEOKUL, W. (2000) . Breaking the silence: setting realistic priorities for AIDS control in less developed countries. Lancet 356, 55-60.

British HIV association (BHIVA) Guidelines Coordinating Committee. (1997). Guidelines for antiretroviral treatment of HIV seropositive individuals. Lancet, 349, 1046-1092.

BLOWER, S.M., GERSHENGORN, H.B., GRANT, R.M. (2000). A tale of two futures : HIV and antiretroviral therapy in San Francisco. Science, 287, 650-654.

CARPENTER, C.C.J., COOPER, D.A., FISCHL, M.A. et al. (1997) Antiretroviral therapy for HIV infection in 1997. Updated recommendations of the International AIDS Society-USA panel. JAMA, 277, 1962-1969.

Center for Diseases Control (CDC). (1998). Report of the NIH Panel to define principles of therapy of HIV infection. Guidelines for the use of antiretroviral agents in HIV-infected adults and adolescents. MMWR , 47 (RR-5), 43-82.

CREESE, A., FLOYD, K., ALBAN, A., AND GUINNESS, L. (2002). Cost-effectiveness of HIV/AIDS interventions in Africa: a systematic review of the evidence. Lancet 359, 1635-1642.

DELFRAISSY, J.F. (Ed.). (1999). Guidelines for the use of antiretroviral therapies in HIV infection. Report to the Ministry of Health & Social Affairs. Flammarion Eds, Paris.

DESCAMPS, D., FLANDRE, P., CALVEZ, V. (2000). Mechanisms of virologic failure in previously untreated HIV-infected patients from a trial of induction-maintenance therapy. Trilège (Agence Nationale de Recherches sur le SIDA 072) Study Team. JAMA, 283, 205-11.

DEEKS, S.G., SMITH, M., HOLODNIY, M. & KAHN, J.O. (1997). HIV-1 protease inhibitors. A review for clinicians. JAMA, 277, 145-153.

DELFRAISSY J. F. (éd.) (2002) Prise en charge des personnes infectées par le VIH. Rapport 2002. Recommandation du groupe d'experts au Ministre de la Santé, de la Famille et des personnes handicapées. Médecine Sciences, Flammarion: Paris.

DIXIT, A. and PINDYCK, R. S. (1994). Investment under Uncertainty. Princeton University Press: Princeton.

FINZI D., BLANKSON J., SILLCIANO JD et al. (1999). 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.

GALLANT, J.E. (2000). Strategies for long-term success in the treatment of HIV infection. JAMA, 283, 1329-1334.

GEOFFARD, P.Y., PHILIPSON, T. (1996). Rational epidemics and their public control. Intl Eco Rev, 37, 603-623.

GERBERT, B., BRONSTONE, A., CLANON, K., ABERCROMBIE, P., BANGS-

BERG, D. (2000). Combination antiretroviral therapy : health care providers confront emerging dilemmas. AIDS CARE, 12, 409-421.

HAMMER, S.M. et al. (2002). Scaling up antiretroviral therapy in resource-limited settings: guidelines for a public health approach. World Health Organization, Geneva.

HECHT, F. M., GRANT, R. M., PETROPOULOS, C. J., et al. (1998). Sexual transmission of an HIV-1 variant resistant to multiple reverse-transcriptase and protease inhibitors. New England Journal of Medicine, 339, 307-11.

HOGG, R.S., HEATH, K.V., YIP, B. et al. (1998a). Improved survival among HIV-infected individuals following initiation of antiretroviral therapy. JAMA, 279, 450-454.

HOGG, R. S.; YIP, B.; CHAN, K. J.; WOOD, E.; CRAIB, K. J.; O'SHAUGHNESSY, M. V.; MONTANER, J. S. (2001). Rates of disease progression by baseline CD4 cell count and viral load after initiating triple-drug therapy, JAMA: The Journal Of The American Medical Association, 286(20), 2568-2577.

IDEMYOR, V. (2002). Continuing debate over HIV therapy initiation, HIV Clinical Trials, 3(2), 173-176.

KITAHATA, M.M., VAN ROMPAEY, S.E., SHIELDS, A.W. (2000). Physician experience in the care of HIV-infected persons is related with earlier adoption of new antiretroviral therapy. J Acquir Immune Defic Syndr , 24, 106-114.

KYRIAKIDES, T. C.; GUARINO, P. (2001). Timing of antiretroviral treatment initiation, JAMA: The Journal Of The American Medical Association, 285(13), 1702-1703.

LANDMAN, R., MOATTI, J.P., PERRIN, V., HUARD, P. & The PAMPA Study Group. (2000). Variability of attitudes toward ealy initiation of HAAT for HIV infection; a study of French prescribing physicians. AIDS CARE, 12, in press.

LAURENT, C. et al. (2002). The Senegalese government's highly active antiretroviral therapy initiative: an 18-month follow-up study. AIDS 16, 1363-1370.

LAXMINARAYAN, R. (2002). Battling Resistance to Antibiotics and Pesticides

Resources for the Future, Washington, D.C..

LE MOING, V.; CHÈNE, G., CARRIERI, M. P; ALIOUM, A.; BRUN-VÉZINET, F.; PIROTH, L.; CASSUTO, J. P.; MOATTI, J.-P.; RAFFI, F.; LEPORT et al. (2002). Predictors of virological rebound in HIV-1-infected patients initiating a protease inhibitor-containing regimen, AIDS (London, England), 16(1), 21-29.

LITTLE, S. J.; HOLTE, S.; ROUTY, J.-P.; DAAR, E. S.; MARKOWITZ, M.; COLLIER, A. C.; KOUP, R.A.; MELLORS, J. W.; CONNICK, E.; CONWAY ET AL. (2002). Antiretroviral-drug resistance among patients recently infected with HIV, The New England Journal Of Medicine, 347(6), 385-394.

MARSEILLE, E., HOFMANN, P.B., AND KAHN, K.G. (2002). HIV prevention before HAART in sub-Saharan Africa. Lancet 359, 1851-1856.

MURPHY, E. L.; COLLIER, A C.; KALISH, L A; ASSMANN, S F; PARA, M F; FLANIGAN, T P; KUMAR, P N; MINTZ, L; WALLACH, F R; NEMO **XX** ET ALXX (2001). Highly active antiretroviral therapy decreases mortality and morbidity in patients with advanced HIV disease, Annals Of Internal Medicine, 135(1), 17-26.

OBADIA,Y., SOUVILLE,M., MORIN, M., MOATTI, J.P. (1999). French general practicioners'attitudes toward therapeutic advances in HIV care : results of a national survey. Intl J STD & AIDS, 10, 243-249.

PATERSON, D. L., SWINDELLS, S., MOHR, J. et al. (2000). Adherence to protease inhibitor therapy and outcomes in patients with HIV infection. Ann Intern Med, 133, 21-30.

PEDRAZA, M. A., DEL ROMERO, J., ROLDAN, F. (1999). Heterosexual transmission of HIV-1 is associated with high plasma viral load levels and a positive viral isolation in the infected partner, J Acquir Immune Defic Syndr, 21, 120-125.

PHILLIPS, A. N.; STASZEWSKI, S.; WEBER, R.; KIRK, O.; FRANCIOLI, P.; MILLER, V.; VERNAZZA, P.; LUNDGREN, J. D.; LEDERGERBER, B.; EUROSIDA STUDY GROUP (2001) HIV viral load response to antiretroviral therapy according to the baseline CD4 cell count and viral load, JAMA: The Journal Of The American

Medical Association, 286(20), 2560-2567.

QUINN, T.C., WAWER, M.J., SEWANKAMBO, N., et al. (2000). Viral load and heterosexual transmission of human immunodeficiency virus type 1. Rakai Project Study Group, N Engl J Med, 342, 921-929.

Recommendations of the Panel on Clinical Practices for Treatment of HIV (2002). Guidelines for using antiretroviral agents among HIV-infected adults and adolescents. Recommendations And Reports: Morbidity And Mortality Weekly Report. Centers For Disease Control, 51(RR-7), 1-55.

REEDJIK, M., BINDELS, P.J.E., MOHRS, J., WIGERSMA, L. (1999). Changing attitudes towards antiretroviral treatment of HIV infection : a prospective study in a sample of Dutch general practicioners. AIDS CARE, 11, 141-145.

RISTIG, M B; ARENS, M Q; KENNEDY, M; POWDERLY, W; TEBAS, P. (2002). Increasing prevalence of resistance mutations in antiretroviral-naïve individuals with established HIV-1 infection from 1996-2001 in St. Louis, HIV Clinical Trials, 3(2), 155-160.

ROSS, S. (1983). Introduction to stochastic dynamic programming. Academic Press, New-York.

RUBIO, R.; BERENGUER, J.; MIRO, J. M.; ANTELA, A.; IRIBARREN, J. A.; GONZALEZ, J.; GUERRA, L.; MORENO, S.; ARRIZABALAGA, J.; CLOTET ET AL. (2002). Recommendations of the Spanish AIDS Study Group (GESIDA) and the National Aids Plan (PNS) for antiretroviral treatment in adult patients with human immunodeficiency virus infection in 2002, Enfermedades Infecciosas y Microbiologia Clinica, 20(6), 244-303.

SCHACKMAN, B. R.; GOLDIE, S. J.; WEINSTEIN, M. C.; LOSINA, E.; ZHANG, H.; FREEDBERG, K. A. (2001). Cost-effectiveness of earlier initiation of antiretroviral therapy for uninsured HIV-infected adults, American Journal Of Public Health, 91(9), 1456-1463.

SENAK, M. (1997) Predicting antiviral compliance: physician's responsibilities vs.

Patients' rights. J Intl Assn Phys in AIDS Care, 3, 45-8.

SILVEIRA, M. P. T.; DRASCHLER, M. DE L.; LEITE, J. C. DE C.; PINHEIRO, C. A. T.; DA SILVEIRA, V. (2002) Predictors of undetectable plasma viral load in HIV-positive adults receiving antiretroviral therapy in Southern Brazil, The Brazilian Journal Of Infectious Diseases: An Official Publication Of The Brazilian Society Of Infectious Diseases, 6(4), 164-171.

SONNABEND, J. The debate on HIV in Africa. Lancet 355, 2163 (2000).

STEWART, G. (1997). Adherence to antiretroviral therapies. In Van PRAAG, E., FERNYAK, S. & KATZ, A.M. The implications of antiretroviral treatments. Informal consultation. Geneva, WHO/UNAIDS, 35-50.

STOKEY, N.L. (1989). Recursive methods in economic dynamic. Harvard University Press, Boston.

TANURI, A.; CARIDEA, E.; DANTAS, M. C.; MORGADO, M. G.; MELLO, D. L. C.; BORGES, S.; TAVARES, M.; FERREIRA, S. B.; SANTORO-LOPES, G.; MARTINS et al. (2002). Prevalence of mutations related to HIV-1 antiretroviral resistance in Brazilian patients failing HAART, Journal Of Clinical Virology, 25(1), 39-46.

TEBAS, P.; HENRY, K.; NEASE, R.; MURPHY, R.; PHAIR, J.; POWDERLY, W. G. (2001). Timing of antiretroviral therapy: Use of Markov modeling and decision analysis to evaluate the long-term implications of therapy, AIDS (London, England), 15(5), 591-599.

VAN HEESWIJK, R. P.; VELDKAMP, A.; MULDER, J. W.; MEENHORST, P. L.; LANGE, J. M.; BEIJNEN, J. H.; HOETELMANS, R. M. (2001). Combination of protease inhibitors for the treatment of HIV-1-infected patients: a review of pharmacokinetics and clinical experience, Antiviral Therapy, 6(4), 201-229.

VAN PRAAG, E., FERNYAK, S., KATZ, A.M. (1997). Impact of antiretroviral treatments. Informal consultation. WHO/ASD/97.2, Geneva.

VOELKER, R. (1997) Debating dual AIDS guidelines. JAMA, 278, 613.

WAINBERG, M.A. & FRIEDLAND, G. (1998). Public health implications of antiretroviral therapy and HIV drug resistance. JAMA, 279, 1977-1983.

WEIDLE, P.J. et al. (2002). Assessment of a pilot antiretroviral drug therapy programme in Uganda: patients' response, survival, and drug resistance. Lancet 360, 34-40.

WOOD, E., BRAITSTEIN, P., MONTANER, J.S.G. et al. (2000). Extent to which low-level use of antiretroviral treatment could curb the AIDS epidemic in sub-Saharan Africa. Lancet, 355, 2095-2100.

YENI, P. G; HAMMER, S. M.; CARPENTER, C. C. J.; COOPER, D. A; FISCHL, M. A.; GATELL, J. M; GAZZARD, B. G; HIRSCH, M. S; JACOBSEN, D. M; KATZENSTEIN et al.. (2002) . Antiretroviral treatment for adult HIV infection in 2002: updated recommendations of the International AIDS Society-USA Panel, JAMA: The Journal Of The American Medical Association, 288(2), 222-235.

ZHANG L., RAMRATNAM B., TENNER-RAQ K et al. (1999). Quantifying residual HIV-1 replication in patients receiving combination antiretroviral therapy. N Engl J Med., ,340, 1605-1613.

Appendix

1.1 The solution of (5)

The first-order derivatives in (5) are:

$$\begin{aligned} \frac{\partial V}{\partial x} &= Ka - Ca^2x - Ca(1-a)y \\ \frac{\partial V}{\partial y} &= J(1-a) - C(1-a)^2 y - Ca(1-a)x. \end{aligned}$$

An interior solution with $x, y \in [0,1]$ requires

$$\begin{aligned} x &= \frac{Ka - Ca(1-a)y}{Ca^2} = \frac{K - C(1-a)y}{Ca} \\ & \text{and} \\ x &= \frac{J(1-a) - C(1-a)^2 y}{Ca(1-a)} = \frac{J - C(1-a)y}{Ca}.f \end{aligned}$$

This implies $K = J$ i.e. (6).

It follows that at least one of x or y is a corner solution, i.e. has a value or 0 or 1.

There are two possible *ex post* therapeutic situations:[15]

- $K > J$ i.e. $\Gamma\Delta(b,G) + (1-\Gamma)\Delta(b,B) - \Delta(b,0) > \Gamma\Delta(0,G) + (1-\Gamma)\Delta(0,B) - \Delta(0,0)$: the expected utility gain in period 2 is higher, relative to no-therapy, for patients who have undergone Therapy 1 in period 1 than for patients who have not undergone any therapy. Thus if one at least of $x \equiv A(b,G)$ or $y \equiv A(0,G)$ is set at 1, it must be x, thereby selecting all patients who have undergone Therapy 1 as candidates for Therapy 2. This is optimal if $\frac{\partial V(1,y)}{\partial x} \geq 0$.[16] Alternativey, if neither x nor y is set at 1, one at least, y, must be set at 0. This is optimal if $\frac{\partial V(x,0)}{\partial y} < 0$.[17] In the first instance the solution is found by setting $x = 1$ and then solving for y; in the second instance the solution is found by setting $y = 0$ and

[15] These therapeutic situations are called *ex post* because they are conditional on Therapy 1 having proven 'bad'.

[16] Since $\frac{\partial^2 J}{\partial x \partial y} = -c^2 a(1-a)$, the condition $\frac{\partial J(1,y)}{\partial x} > 0$ is met for all y if it is true at $y = 1$.

[17] Since $\frac{\partial^2 J(x,0)}{\partial y \partial x} = -c^2 a^2$, this condition is met for all x if it is true at $x = 1$.

then solving for x. Thus:

$$\begin{aligned}
\text{if } K &> J \text{ and } \frac{\partial V(1,y)}{\partial x} \geq 0, \text{ then} \\
x &= 1 \text{ and} \\
y &= \begin{cases} 1 \text{ if } \frac{\partial V(1,1)}{\partial y} \geq 0, \\ \frac{J-aC}{(1-a)C} \text{ if } \frac{\partial V(1,y)}{\partial y} = 0, \\ 0 \text{ if } \frac{\partial V(1,0)}{\partial y} < 0. \end{cases}
\end{aligned} \tag{A.1}$$

Conversely[18]

$$\begin{aligned}
\text{if } K &> J \text{ and } \frac{\partial V(x,0)}{\partial y} < 0, \text{ then} \\
y &= 0 \text{ and} \\
x &= \begin{cases} 1 \text{ if } \frac{\partial V(1,0)}{\partial x} \geq 0, \\ \frac{K}{aC} \text{ if } \frac{\partial V(x,0)}{\partial x} = 0 \end{cases}
\end{aligned} \tag{A.2}$$

The alternative therapeutic situation is when:

- $K < J$ i.e. $\Gamma\Delta(b,G) + (1-\Gamma)\Delta(b,B) - \Delta(b,0) < \Gamma\rho\Delta(b,G) + (1-\Gamma)\Delta(0,B) - \Delta(0,0)$: the expected utility gain is higher, relative to no-therapy, for patients not having undergone any therapy in period 1. Following the same argumentation as above, if one of $x \equiv A(b,G)$ or $y \equiv A(0,G)$ is set at 1, it must be y, thereby selecting all patients who have not undergone any therapy 1 as candidates for Therapy 2. This is optimal if $\frac{\partial V(x,1)}{\partial y} \geq 0$.[19] Alternativey, if neither x nor y is set at 1, one at least must be set at 0. The least attractive category, x, is thus left untreated, which is optimal if $\frac{\partial V(0,y)}{\partial x} < 0$.[20] In the first instance the solution is found by setting $y = 1$ and then solving for x; in the second instance the solution

[18] Since $a \leq 1$, both conditions can be met simultaneously. Also, if $y = 0$, 0 it is impossible for $\frac{\partial J(0,0)}{\partial x}$ to be negative, i.e. $Ka \leq 0$.

[19] Since $\frac{\partial^2 J}{\partial x \partial y} = -c^2 a(1-a)$, this condition is met for all x if it is true at $x = 1$.

[20] Since $\frac{\partial^2 J(x,0)}{\partial y \partial x} = -c^2 a^2$, this condition is met for all y if it is true at $y = 1$.

is found by setting $x = 0$ and then solving for y. Thus:

$$\begin{aligned} \text{if } K &< J \text{ and } \frac{\partial V(x,1)}{\partial y} \geq 0, \text{ then} \\ y &= 1 \text{ and} \\ x &= \begin{cases} 1 \text{ if } \frac{\partial V(1,1)}{\partial x} \geq 0, \\ \frac{K-(1-a)C}{aC} \text{ if } \frac{\partial V(x,1)}{\partial x} = 0, \\ 0 \text{ if } \frac{\partial V(0,1)}{\partial x} < 0. \end{cases} \end{aligned} \tag{A.3}$$

Conversely[21]

$$\begin{aligned} \text{if } K &< J \text{ and } \frac{\partial V(0,y)}{\partial x} < 0, \text{ then} \\ x &= 0 \text{ and} \\ y &= \begin{cases} 1 \text{ if } \frac{\partial V(0,1)}{\partial y} \geq 0, \\ \frac{J}{(1-a)C} \text{ if } \frac{\partial V(0,y)}{\partial y} = 0. \end{cases} \end{aligned} \tag{A.4}$$

The above cases are mutually exclusive and cover all possibilities. Treating a as a parameter at this stage, the parameter constraints derived from the conditions on $\frac{\partial V(x,y)}{\partial x}$ and $\frac{\partial V(x,y)}{\partial y}$ that arise in each case, and from the conditions that both x and y belong to the interval $[0,1]$, are gathered in Table 1, corresponding to $K < J$, given in the text, and Table A, for $J < K$, given below for completeness.

Table A: Optimal period 2 shares and net utility when Therapy 1 proves ineffective while it was more attractive than no therapy *ex ante* $(K > J)$

Param. cond's	Optimal period 2 shares (x^*, y^*)	locus	Optimized Value $V^*(a\|b)$
$C \leq J < K$	$(1,1)$	A_b	$a(K-J) + J + \Delta - \frac{1}{2}C$
$aC < J < C$	$\left(1,\ 0 < \frac{J-aC}{(1-a)C} < 1\right)$	B'_b	$a(K-J) + \Delta + \frac{1}{2}\frac{J^2}{C}$
$J \leq aC \leq K$	$(1,0)$	C'_b	$aK + \Delta - \frac{1}{2}Ca^2$
$J < K < aC$	$\left(0 < \frac{K}{aC} < 1,\ 0\right)$	D'_b	$\frac{1}{2}\frac{K^2}{C} + \Delta$

[21] Since $a \leq 1$, both conditions can be met simultaneously. Also, if $y = 0$, 0 it is impossible for $\frac{\partial J(0,0)}{\partial x}$ to be negative, i.e. $Ka \leq 0$.

1.2 Condition for an interior period 1 solution ($\frac{\partial W}{\partial a} = 0$)

Considering (8) , (9) , and Table 2, we have $\frac{\partial W(a)}{\partial a} =$

$$\begin{cases} -ca + \gamma\delta(g) + (1-\gamma)\delta(b) - \delta(0) + \gamma(M-J) & \text{if} \quad 0 \leq a \leq 1 - \frac{J}{C} \\ -(c + (1-\gamma)C)a + \gamma\delta(g) + (1-\gamma)(\delta(b) + C) - \delta(0) + \gamma M - J & \text{if} \quad 1 - \frac{J}{C} < a \leq 1 - \frac{K}{C} \\ -ca + \gamma\delta(g) + (1-\gamma)\delta(b) - \delta(0) + K + \gamma(M-K) & \text{if} \quad 1 - \frac{K}{C} < a \leq \frac{J}{C} \\ -(c + \gamma C)a + \gamma\delta(g) + (1-\gamma)\delta(b) - \delta(0) - (1-\gamma)(J-K) + \gamma M & \text{if} \quad \frac{J}{C} < a \leq 1 \end{cases}$$

The optimum value of a is interior if the value obtained by setting $\frac{\partial W(a)}{\partial a} = 0$ also belongs to the interval of validity of that derivative, i.e.:[22]

[22] The four lines of (10) correspond to the last four lines of Table 2.