European Journal of Operational Research 201 (2010) 308–318

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Innovative Applications of O.R.

When in a drug epidemic should the policy objective switch from use reduction to harm reduction?

Jonathan P. Caulkins [a,b,*], Gustav Feichtinger [c], Gernot Tragler [c], Dagmar Wallner [c]

[a] Carnegie Mellon University, H. John Heinz III School of Public Policy Management, 5000 Forbes Ave., Pittsburgh, PA 15213, USA
[b] Carnegie Mellon University in Qatar, PO Box 24866, Doha, Qatar
[c] Institute for Mathematical Methods in Economics, Vienna University of Technology, Argentinierstrasse 8/105-4, A–1040 Vienna, Austria

ARTICLE INFO

Article history:
Received 19 July 2007
Accepted 6 March 2009
Available online 25 March 2009

Keywords:
OR in societal problem analysis
Control
Health
Drug policy
Epidemics

ABSTRACT

A heated debate in drug policy concerns the relative merits of "harm reduction" (e.g., reducing drug-related HIV/AIDS transmission) vs. "use reduction" (controlling drug use per se). This paper models whether shifting emphasis between these goals over the course of a drug epidemic might reduce social costs relative to pursuing one or the other exclusively. Results suggest different answers for different drugs and/or countries. In particular, harm reduction may have always been effective for Australia's injection drug use problem, but for US cocaine it may not have been in the past even if it could be so today. In certain circumstances harm reduction may "tip" an epidemic toward a high- rather than low-use equilibrium. The location in state space of regions where this occurs can be sensitive to parameter changes, suggesting caution may be in order when advocating harm reduction, unless there is confidence the epidemic has been modeled and parameterized accurately.

1. Introduction

A vigorous debate rankles within drug policy between those advocating "harm reduction" and those placing primacy on reducing drug use, referred to here as "use reduction" (cf., Caulkins and Reuter, 1997). Previous work (Behrens et al., 1999, 2000; Tragler et al., 2001; Winkler et al., 2004) suggested a peaceful resolution to the equally contentious debate between proponents of "demand reduction" and "supply control", namely, that each had a crucial role to play, but their relative effectiveness could vary over an "epidemic cycle" of drug abuse. This paper explores the possibility of a similar resolution to the harm vs. use reduction debate.

The term "harm reduction" is politically charged. While "harm reduction" is official policy in several countries (notably, Australia, the Netherlands, and the United Kingdom), it is denounced by national drug policy makers in the United States as a deceitful ploy used by covert advocates of legalization (e.g., McCaffrey, 1998). To complicate matters, the term is used to mean different things by different people.

One definition is essentially synonymous with minimizing social costs. MacCoun (1998) describes this as "macro harm reduction" to distinguish it from "micro harm reduction" meaning minimizing how harmful drugs are, particularly to users. The distinction is captured in the simple equation:

$$\text{Total harm} = \text{Total use} * \text{average harm per unit of use}. \quad (1)$$

Macro harm reduction means reducing total harm, which is the left side of Eq. (1) and will be our objective function below. Micro harm reduction stands for reducing the far right hand term of Eq. (1). This reduction of the average harm attributable to a unit of consumption is what we mean when using the term "harm reduction" here.

Advocates of "use reduction" try to reduce total harm by reducing use, either viewing harmfulness as immutable or worrying that efforts to reduce harmfulness will spur greater use. Critics of use reduction argue it is hard to suppress drug use in a free society and/or that efforts to suppress use displace it into more harmful forms. The classic example offered is that prohibiting syringes leads injection drug users to share and reuse syringes, exacerbating the spread of blood-borne diseases, notably HIV/AIDS and Hepatitis C. They suggest it is more productive to make use safer, e.g., by providing supervised injection facilities (SIFs) where overdoses can be quickly detected and treated. They downplay fears that such measures "send the wrong message" or induce more youth to try drugs (Reuter and Caulkins, 1995).

What the equation does not fully specify is whose harms get counted. Many harm reduction advocates focus on harms suffered by users, notably the risk of overdose and HIV/AIDS. Many advocates of use reduction focus on harms to non-users, notably drug-related crime and violence. Here, we parameterize the

* Corresponding author. Address: Carnegie Mellon University, H. John Heinz III School of Public Policy Management, 5000 Forbes Ave., Pittsburgh, PA 15213, USA. Tel.: +1 412 268 9590; fax: +1 412 268 5338.
E-mail addresses: caulkins@andrew.cmu.edu, caulkins@cmu.edu (J.P. Caulkins), or@server.eos.tuwien.ac.at (G. Feichtinger), tragler@server.eos.tuwien.ac.at (G. Tragler), dagmar.wallner@gmail.com (D. Wallner).

0377-2217/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.ejor.2009.03.015

objective function in terms of social costs reflected in so-called "Cost of Illness" (COI) studies (Harwood et al., 1998; Collins and Lapsley, 2002), which include crime costs. COI studies have limitations (Reuter, 1999; Kleiman, 1999; Moore and Caulkins, 2006), but they remain the de facto standard. If and when better social cost estimates are produced, the model could just as easily be parameterized with them.

It is not surprising that Australia and some European countries differ from the US with respect to harm vs. use reduction objectives. In the US, crime and violence dominate social costs and public concern about illicit drugs, and two-thirds of those costs stem from a drug (cocaine, including crack) that is not usually injected in the US (Caulkins et al., 2002). In most other developed countries, violence is much less of a concern and the dominant drug is heroin (UNODC, 2005), which is usually injected and which has a particularly high risk of overdose because of its low "safety ratio" – the ratio of the usual lethal dose to the usual recreational dose (Gable, 2004).

We explore the possibility that both policy regimes may be preferred, but in different places and/or times. In particular, we parameterize a dynamical systems model of drug use for both Australian injection drug use (IDU) and US cocaine use, to see whether the drugs and/or national contexts influence the potential benefits of harm reduction.

We run the model as a policy simulation to see whether the policy that is best also depends on the stage of the epidemic. This concept of "stage of the epidemic" merits explanation because it also is controversial. There is of course no pathogen that infects users with drug use, the way that pathogens spread malaria, the flu, or HIV. However, drug use is "contagious" in the same way that fashions, laughter, and even rumors can be (Noymer, 2001). Almost everyone who starts using illicit drugs is introduced to them by a friend, sibling, or acquaintance. The myth of a "drug pusher" seducing unsuspecting naifs is just that, a myth (Coomber, 2006). Just as the marketing literature has built new product adoption models in which first time use of a consumer electronics gadget is driven by how many people are already using that gadget (e.g., Bass, 1969), so too can one model the spread of drug use with such models (Caulkins, 2005).

Some people hate the term "epidemic" even in this metaphorical sense because they fear it will stigmatize users and/or justify draconian and inhumane measures along the lines of quarantines. We recognize that risk, but nevertheless prefer to use the term because what happens with drug use is not just some general spread ("diffusion") but specifically a process of social interaction between users and non-users that can lead the non-user to begin using (cf., Ferrence, 2001).

This contagious spread can lead to explosive growth in drug use followed by a peak and decline. Many models of drug use and drug markets have a "tipping point" (Baveja et al., 1993; Caulkins, 1993; Tragler et al., 2001). When few people are using or selling the drug, it is relatively easy to keep the drug from spreading for several reasons. In a "thin" market, sellers and buyers have a hard time locating each other; behaviors that are uncommon are more likely to be socially stigmatized; and/or a modest level of enforcement effort can create high risks for each of the relatively small number of people over whom that effort is distributed (Kleiman, 1993). At the other extreme, when the drug is very widely used, having one more or one fewer person use might have very little effect on the likelihood of others trying the drug because potential users will be offered the drug multiple times even without that additional user and, if they chose to use, will be able to locate multiple suppliers (Riley, 1997). In between, however, there may be a tipping point where the market is still of modest size, but close to reaching a critical mass that will enable it to spread widely. For markets near that tipping point, small changes can have large and lasting effects on the long-run trajectory of the prevalence of use.

Given this overview of epidemic concepts, let us return to the question of whether harm reduction is good policy or not. If harm reduction truly has no adverse effect on how many people use drugs, then the question is trivial; reducing harmfulness would clearly reduce total harm (cf., Nordt and Stohler, 2006). However, one does not have to believe people have perfect rationality to worry that more people might participate in an activity if that activity is made safer. MacCoun (1998) reviews psychological and empirical evidence concerning this possibility in a variety of domains. He concludes that there can be "risk compensation", so making an activity safer generally increases participation in that activity. As examples, he cites literature finding that "drivers have responded to seat belts and other improvements in the safety of automobiles by driving faster and more recklessly" and "filters and low-tar tobacco each reduce the harmfulness per unit of tobacco, yet numerous studies have demonstrated that smokers compensate by smoking more cigarettes, inhaling more deeply, or blocking the filter vents" (p. 1203). A more recent concern is the possibility that HIV vaccines and/or effective AIDS treatment could reverse some of the risk reducing behaviors adopted in response to the HIV/AIDS epidemic (Blower et al., 2002).

Risk compensation means that reducing harmfulness has ambiguous effects on total harm. If participation went up enough, total harm could go up even if harm per unit of activity went down. However, MacCoun argues there is little evidence of compensatory responses producing greater than proportionate changes in individual behavior. Consistent with this, we employ a model of risk compensation in which the percentage change in initiation into drug use is always less than the percentage change in risk to the user. Furthermore, we assume harm reduction has no adverse effect on exit from drug use.

However, as mentioned, drug use is characterized by feedback effects, specifically feedback from current use to initiation. Hence, a possibility that remains to be explored is whether such nonlinearity means micro harm reduction could increase macro harm even when that could not happen if there were no feedback.

In particular, if harm reduction does have some perverse effect on initiation and drug epidemics have tipping points, then it seems plausible that a harm reduction policy could be a bad idea at some points (specifically when the epidemic is near its tipping point) even if it is a good idea at other times (after the drug is already in pervasive use and perhaps when it is so rare that there is little risk of use exploding).

Testing this conjecture empirically is difficult. There are very few jurisdictions (none?) for which there are good long-term data on drug use and drug-related social costs and which have switched from a use reduction to a harm reduction policy or vice versa. In the absence of empirical data, it can be useful to construct a model to lend precision to an idea and expose inconsistencies that may lurk beneath the imprecision inherent in arguments couched in natural language. The next section introduces such a model. It is an extremely simple model to reduce the risk that the results emerge only from certain atypical structures and/or that the results cannot be interpreted intuitively.

2. Conceptual model

We take as our point of departure the classic Bass model (Bass, 1969). It posits that the rate of adoption by non-users who are susceptible to use is linearly increasing in the number of current users, expressed here as a proportion of the total population and denoted by $A(t)$. In the original Bass model, product adoption spreads quickly relative to life spans, so the number of people susceptible

to initiation is just the total number who were originally susceptible minus the number who have already adopted. Hence, the term modeling adoption is

$$\text{new adoptions per unit time} = (i_1 + i_2 A(t))(1 - A(t)). \quad (2)$$

In the present context a more sophisticated tracking of susceptibles is warranted, so we capture their numbers with a second state variable, denoted $S(t)$ and interpret $A(t)$ as the actual number, not the proportion who have adopted. Given this modification, the number of new adoptions per unit time is:

$$\#\text{ of new adoptions per unit time} = (i_1 + i_2 A(t))S(t). \quad (3)$$

The coefficient i_1 is referred to as the coefficient of innovation because it reflects adoption that is independent of the current level of use. Coefficient i_2 is referred to as the coefficient of imitation because it reflects adoption stemming from contact with a current user. Note: drug use prevalence and drug-related rates are low enough that the drug epidemic is not a major driver of the country's total population. That means that the proportion of people who are drug users is essentially $A(t)$ divided by a constant, and so is proportional to $A(t)$. Hence, if one imagines the user and susceptible populations mix randomly, then the number of contacts between users and non-users is proportional to the product, $A(t)S(t)$. That means i_2 can be seen as being proportional to the "infectivity" or probability that such an interaction leads to a new "infection".

Assuming i_2 is constant makes sense in many contexts. When a person with the flu sneezes on someone who is not yet sick, the probability of infection may depend on a lot of factors (e.g., age, health-status, whether the person got a flu vaccine), but not on how many other people do or do not have the flu.

However, when becoming "infected" is the result of a conscious choice, the virulence of the contagion can depend on the prevalence of the epidemic. So when someone who is infected (a current user) interacts with a "susceptible" (a current non-user) the probability that the "susceptible" becomes infected need not be a constant. Rather, that probability of infection per interaction may be a function of the prevalence.

Consider, for example, the application of such models to fashion goods. When a consumer good is widely used, that can make the good either more or less appealing to others. The appeal of certain high-end fashion goods stems from their exclusivity or snob effect. Fashion-conscious consumers pay high prices for designer brand items in part because they know, and others watching them know, that the masses cannot afford to purchase such exclusive goods. So, high prevalence would reduce infectivity. On the other hand, there are popular fashions that become more not less appealing when they have already been adopted by others, particularly for teenagers. Furthermore, Bundgaard-Neilsen (1976) has argued that late adopters may adopt faster than earlier ones because they are in a better position to assess the new technology. So adoption rates might be either concave or convex in the number of users. Indeed, the Non-Uniform Influence Diffusion Model by Easingwood et al. (1983) extends the original Bass model based on such ideas. In a one-state framework, they assess the form

$$\#\text{ of adoptions per unit time} = (i_1 + i_2 A(t)^{\delta})(1 - A(t)). \quad (4)$$

The authors illustrate the generality of the model for five goods. The empirical study leads to $\delta < 1$ in four of the cases, whereas in one case the authors find $\delta > 1$. In our two state framework, the Bass model is generalized by using a function $g(A(t))$ modeling dependence of adoption on the current number of users. We allow users to exit the population (by death or ceasing use) at a constant per capita rate, μ. (Since most drug users are young and only a subset are dependent users with substantially elevated death risk, most exit from the drug using population is through quitting, not death.) This leads to

$$\#\text{ of new adoptions per unit time} = (i_1 + i_2 g(A(t)))S(t), \quad (5)$$

$$\#\text{ of quitters} = \mu A(t), \quad (6)$$

where we presume $g(0) = 0$ since there should not be any imitation if there is no one to imitate. For notational simplicity we will let $f(A(t))$ stand for the entire term $(i_1 + i_2 g(A(t)))$. We will pay particular attention to the case when $g(A(t))$ is a simple power function of the form $A(t)^{\alpha}$, including the special case when i_1 is negligible.

We expect that $g()$ is an increasing function of $A(t)$, the current prevalence of use, but it may be either concave or convex, as in the high- and low-fashion examples mentioned above, respectively.

In many respects illicit drugs are a popular fashion. When someone is offered the chance to use an illicit substance for the first time, they might be less likely to accept if little is known about the drug and/or its use is highly deviant in the sense of being confined to small, highly atypical populations, but more likely to accept the offer if many of their peers are already using it. (Most drug initiation occurs among teens and young adults for whom the desire for conforming with peers' lifestyles can be a powerful motivator.) Hence, we might expect $g()$ to be convex for at least some illicit drugs. Indeed, parameterization of the current model for cocaine use in the U.S. involves $\alpha > 1$.

On the other hand, some of the most severe adverse consequences of drug use manifest only later, when use is relatively widespread and some people have been using for an extended period (Musto, 1987). In such cases, virulence may decline as use spreads. An entirely separate argument with the same bottom line is that when use is common, non-users might receive multiple offers, and people who have rejected earlier offers may be less likely to accept subsequent ones. Hence, as use grows, non-users might become "saturated" with opportunities to use, and expansions in numbers of users bring less than proportionate increases in initiation. Indeed, below we will use a concave $g()$ for Australian IDU.

Finally, return to the question of modeling non-users who are susceptible to use. Very few people who are not teen-agers or young adults ever start using drugs, so we can think of the pool of susceptibles as turning over every ten to twenty years. The inflow is people, or a subset of people, who are just coming of age (perhaps turning age 12 if one wants to be concrete). There are two outflows: (1) initiating drug use, which moves individuals into the $A(t)$ population and (2) other exit, which includes death, out migration, etc., but primarily reflects "maturing out" of the pool of susceptibles (again to be concrete, one might think of this as turning 30).

Hence, we let the inflow to $S(t)$ be a simple constant, k, (representing each successive birth cohort as it reaches the age when vulnerability to drug use begins) with a constant per capita outflow rate, δ, that is roughly equal to one over the duration of time during which people are likely to consider starting to use drugs (roughly ten to twenty years).

This discussion motivates the next section's model. Before proceeding, note that there are broadly two kinds of fashions: good (or at least innocuous) fashions, such as particular clothing or music styles, and "bad" fashions, such as drug use, that are socially harmful. Below we consider the case of controlling harms associated with a bad fashion, but one could imagine adapting the basic framework to explore how a firm might promote adoption of its "good" fashion product.

3. Mathematical model

3.1. Formulation

The objective is to minimize total discounted drug-related social costs over some planning horizon by adjusting the extent to which harm reduction is pursued, while drug use evolves

according to epidemic dynamics that may be influenced by harm reduction interventions. The control variable $v(t)$ is the percentage reduction in the harmfulness of drug use, and at any given time, social costs are the product of (1) the number of drug users, $A(t)$, (2) a baseline social cost per user per unit time when there is no harm reduction, which is normalized to 1 without loss of generality, and (3) one minus $v(t)$, the proportion of harm that is not averted via harm reduction policies.

One could penalize control spending, by adding a term $c(v)$ to the objective function, but harm reduction programs receive very modest levels of funding even in countries such as Australia and the Netherlands that make harm reduction the centerpiece of their national policies (Moore, 2005; Rigter, 2006). It is better to think of harm reduction not as a program with a budget but rather as a policy. For example, when a jurisdiction pursues the harm reduction policy of telling police not to arrest people for possessing a syringe, implementing that policy costs essentially nothing more than the paper on which the new policy memos are printed. Consequently and for the sake of simplicity, we omit this $c(v)$.

There are, however, limits on the extent to which a drug's harmfulness can be reduced. Hence, we imposed an upper bound $v(t) \leqslant v_{max}$, whose specific value will vary by drug and country and is discussed further below.

The benefit of control v is obvious; it directly reduces the instantaneous objective function cost. The downside is that it might increase initiation. No one knows for sure that harm reduction has this downside, but it is plausible. Precisely because no one has empirical data on the extent (if any) of this effect, initially we denote it by a general non-decreasing function $h(v)$ with $h(0) = 1$ that multiplies the initiation term. Hence, the general formulation we consider is:

$$\begin{aligned} J &= \int_0^\infty e^{-rt}(A(1-v))dt \\ \dot{S} &= k - \delta S - f(A)Sh(v) \\ \text{s.t.} \\ \dot{A} &= f(A)Sh(v) - \mu A. \end{aligned} \tag{7}$$

Solving the optimal dynamic control version of this problem with v allowed to vary over time (i.e., $v = v(t)$ with $0 \leqslant v(t) \leqslant v_{max}$) would be of interest. However, we consider the simple case of static optimization. That is, we assume implementation of harm reduction is a one-time, irrevocable decision. We focus not on choosing the best value of v for a particular set of initiation conditions, $S(0)$ and $A(0)$, but rather on comparing the performance of pure use reduction ($v(t) = 0$ for all t) vs. pure harm reduction ($v(t) = v_{max}$ for all t) for initial conditions corresponding to various points in a drug epidemic.

3.2. Analysis for general f(A) and h(v)

Some analysis can be done without specifying how the drug's virulence depends on the current prevalence ($f(A)$) or how harm reduction affects initiation ($h(v)$). In particular, when adding the two state equations, the $f(A)$ and $h(v)$ terms cancel, leaving

$$\dot{S} + \dot{A} = k - \delta S - \mu A \tag{8}$$

so the steady state values (denoted by the superscript ^) satisfy

$$\widehat{S} = \frac{k - \mu\widehat{A}}{\delta}. \tag{9}$$

The isocline $\dot{A} = 0$ is given by $\mu\widehat{A} = f(\widehat{A})\widehat{S}h(v)$, which yields the equation

$$\widehat{S} = \frac{\mu\widehat{A}}{f(\widehat{A})h(v)}. \tag{10}$$

In the case without innovators ($i_1 = 0$) and for the power function $g(A) = A^\alpha$, for which $g(0) = 0$, it is easy to see that there is a steady state with no use ($\widehat{A} = 0$), the maximum possible number of susceptibles ($\widehat{S} = k/\delta$), and no control needed, $v = 0$.

More generally, we have two Eqs. (9) and (10) relating the steady-states that are of the form $\widehat{S}(\widehat{A})$. The first one, Eq. (9), is a downward sloping line that is independent of the control v. Since $h(v)$ is increasing in v, exerting the control v pulls down the curve given by (10), shifting steady states to the lower right along the first line. That is, harm reduction leads to more users and fewer susceptibles in steady state – exactly as one would expect.

If $g(A)$ is concave for $A > 0$, then Eq. (10) is strictly increasing in A, so there is exactly one interior steady state (in addition to the steady state $(\widehat{S}, \widehat{A}) = (k/\delta, 0)$), and it is easy to show that interior steady state is stable.

If $g(A)$ is convex, then Eq. (10) is decreasing and convex in A, asymptotically approaching the horizontal A-axis. Furthermore, at $A = 0$, Eq. (10) starts above Eq. (9). Hence, there are three possibilities. Eq. (10) could remain always above Eq. (9) (no interior steady states), be just tangent to Eq. (9) yielding a single interior steady state, or cross it twice, yielding two interior steady states. Appendix B shows that when there are two interior steady states, then from left to right the three steady states (including the one at the origin) are stable, saddle, and stable, respectively.

Since for both the US and Australian parameterizations, i_1 was estimated to be zero, we do not pursue the analysis of $i_1 > 0$ further.

3.3. Specifying functional forms for f(A) and h(v)

To proceed further, we specify functional forms for $f(A)$ and $h(v)$. The parameterizations exhibit $i_1 = 0$. Following Tragler et al. (2001), we assume $g()$ is a power function in A, so $f(A) = i_2 A^\alpha$. This simple form has the flexibility to accommodate both concave and convex functions.

To the best of our knowledge no one has attempted to model explicitly how harm reduction affects use. In all likelihood, the answer is quite complicated, and a psychologically realistic model would consider users' imperfect knowledge, foresight or lack there of, discount rates and form (hyperbolic vs. exponential discounting), over confidence, loss aversion, etc. We take the much simpler approach of imagining that changes in the non-monetary costs of using drugs (e.g., the health risks) affect use the same way as do changes in the monetary costs of using drugs. This is consistent with the so-called "search time" theory of local drug markets (Kleiman, 1988), and it is convenient because there is a growing empirical literature that estimates how responsive drug use is to changes in drug prices (Grossman, 2004).

The personal, non-monetary costs of drug use are conceptually distinct from the social costs of drug use in two ways. First, some social costs are externalities from the users' perspective. Notable among those are the costs drug-related crimes impose on third parties. There is no reason why the government should not pursue policies to reduce the magnitude of these externalities, e.g., by pushing drug markets away from brazen street corner dealing with all its attendant violence and into more covert forms that are less destructive to the community. However, as important as those strategies may be, they are orthogonal to the current model because they do not risk eliciting an adverse behavioral response. Here we focus only on programs that reduce harms felt by users. As a result, the control's upper bound, v_{max}, is the proportion of social costs that are borne by the user. As a proxy, we presume that health-related costs are borne by the user whereas other cost-of-illness (COI) study cost components are externalities. Hence, we set v_{max} to equal the proportion of COI costs that are health-related.

The second distinction is that users may not even factor health-related costs into their consumption decisions as fully as would a

social planner for two reasons. First even some health-related costs may be externalities. For example, when a drug user overdoses, some of the resulting costs are borne by the user (e.g., reduced income because of lost work time due to morbidity or premature mortality), but others are not (e.g., medical treatment covered by Medicaid or emergency care provided at no charge). Second, drug users may be more present-oriented than is a social planner (Kirby and Petry, 2004), so deferred costs of use may not be weighted as heavily as a social planner would like. We let parameter ω denote the proportion of all drug-related health costs that are borne and recognized by the user and in the absence of empirical evidence set it equal to 0.5. Then if c_m and c_s denote the monetary cost and the social cost per unit of consumption, respectively, the cost of use felt by the user without harm reduction is $c_m + \omega v_{max} c_s$, whereas harm reduction reduces that to $c_m + \omega(v_{max} - v)c_s$. So a constant elasticity demand model suggests that

$$h(v) = \left(\frac{c_m + (v_{max} - v)\omega c_s}{c_m + \omega v_{max} c_s}\right)^{\gamma}. \tag{11}$$

Note, we only need to synchronize the units of c_s with c_m with each other, and not with the objective function coefficient (which is normalized to 1), because their units cancel in the $h(v)$ expression.

The exponent γ is the elasticity of participation with respect to the full cost of using. What has been estimated empirically by Dave (2004) and others is the elasticity of participation with respect to the monetary price c_m. Interpreting $h()$ to be a function of c_m, we get

$$h(c_m) = \left(\frac{c_m + (v_{max} - v)\omega c_s}{c_0}\right)^{\gamma}, \tag{12}$$

where c_0 is a constant equal to the baseline cost without harm reduction. Per definition, η is

$$\eta = \frac{dh(c_m)}{dc_m}\frac{c_m}{h(c_m)} = \frac{\gamma c_m}{c_m + (v_{max} - v)\omega c_s}. \tag{13}$$

This expression links our parameter γ to the empirically measured price elasticity of participation by the relation

$$\gamma = \frac{(c_m + (v_{max} - v)\omega c_s)}{c_m}\eta. \tag{14}$$

Note, with the model parameters estimated below, this function $h(v)$ is only modestly convex, almost linear. If a linear approximation were used, the solution to an optimal control formulation that allowed intermediate values of v would nonetheless use bang–bang controls. This suggests that when our analysis below restricts the control to boundary solutions in which harm reduction is pursued aggressively or not at all, that restriction might not matter as much as it would have if the $h(v)$ function were not so nearly linear.

3.4. Parameterization

Table 1 summarizes the parameter values derived in Appendix A for the Australian IDU and US cocaine epidemics.

Structurally the most important parameter is the exponent, α, of the power function $g()$. Since $\alpha > 1$ for US cocaine, $f()$ is convex for $A > 0$ and it is possible to have multiple stable equilibriums separated by a tipping point. In contrast, since $\alpha < 1$ for Australian IDU, there can be only one steady state with a positive amount of drug use, and it is stable. The absence of a tipping point for Australian IDU implies that harm reduction has greater potential to trigger catastrophic increases in use for US cocaine.

The potential benefit of harm reduction is greater for Australian IDU because its upper bound, v_{max}, is much larger (0.53 vs. 0.174 for US cocaine). That reflects the fact that much of the harm associated with Australian IDU comes from outcomes for which effective harm reduction tactics exist (primarily preventing overdose and the spread of blood-borne infectious diseases), whereas more social costs associated with US cocaine pertain to crime, violence, and reduced labor productivity.

4. Results

4.1. Base case results for Australian IDU and US cocaine epidemic models

With both the US cocaine and Australian IDU parameterizations, we computed the net present value of use ($A(t)$) and harm ($A(t)(1 - v(t))$) for various initial numbers of users ($A(0)$), setting the initial number of susceptibles to be the corresponding steady state value ($S(0) = k - \mu A(0)/\delta$). Inasmuch as $A(t)$ tends to be increasing in the early stages of a drug epidemic, this essentially models different points in time at which harm reduction could begin. Fig. 1 shows the results, with $A(0)$ expressed as a proportion of its positive steady state value when there is no harm reduction. This x-axis normalization allows results for both models to be shown on the same graph. (Otherwise numbers of US cocaine users are much larger than are numbers of Australian IDUs.)

For the Australian IDU parameterization, regardless of $A(0)$, harm reduction increases the (present values of) numbers of users but reduces aggregate harm. Unless $A(0)$ is quite small, implementing harm reduction increases drug use by 8–20% and reduces harm by 44–49%.

Results for US cocaine are qualitatively similar for $A(0)$ greater than about 30% of the steady state value, with full harm reduction increasing the present value of future use by 7–12% and reducing corresponding harm by 8–12%. Likewise, at the other extreme, for $A(0)$ less than about 10.8% of the steady state number of cocaine

Table 1
Parameter values.

Parameter	Symbol	Australian IDU	US cocaine
Inflow into S state	k	0.0526	1.3417
Exit rate from S state	δ	0.0952	0.0605
Coefficient of innovation	i_1	0	0
Coefficient of imitation	i_2	0.5112	0.0090
Exponent of $A(t)$	α	0.8622	1.5604
Exit rate from A state	μ	0.1136	0.1661
Social cost of use	c_s	$39,225/yr	$223.56/gm
Proportion of social cost HR can avert	v_{max}	53%	17.408%
Proportion of health costs internalized	ω	0.5	0.5
Monetary cost of use	c_m	$13,537/yr	$106.54/gm
Price elasticity of participation	η	−0.21	−0.45
Cost elasticity of participation	γ	−0.371	−0.532
Annual discount rate	r	0.04	0.04

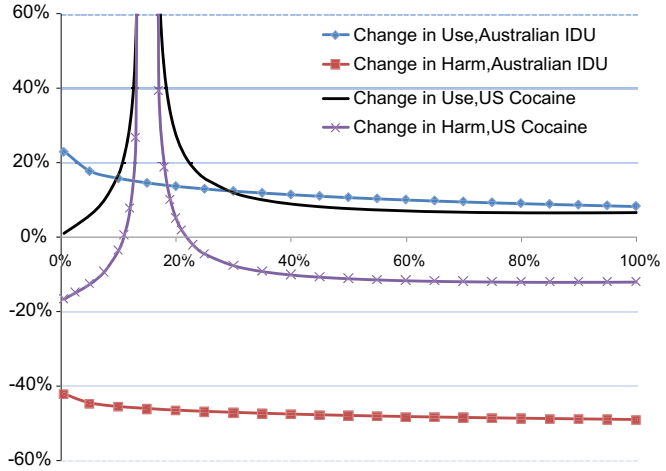

Fig. 1. Comparison of effects of implementing harm reduction with different initial numbers of users for US cocaine and Australian IDU epidemics.

users, harm is reduced even though use goes up. However, for $A(0)$ between 10.9% and 22% of the positive steady state value (roughly 600,000–1,200,000 users), the choice $v(t) = v_{max}$ for all t increases not only drug use but also total harm, dramatically so for certain $A(0)$. For $A(0)$ around 15% of the steady state value, the present value of total future harm can be increased by more than a factor of 4.5.

The reason for these divergent results is that the US cocaine parameters generate a tipping point separating a high-level equilibrium (with $A = 5.49$ million users in the case $v(t) = 0$ for all t and with $A = 5.78$ million when $v(t) = v_{max}$ for all t) from a low-level equilibrium (with $A = 0$ in both cases). With $S(0)$ set as described and no harm reduction, for any $A(0)$ up to about 16% of the high-level steady state (i.e., about 890,000 users), use will ebb back toward the equilibrium with no use. However, with harm reduction, that tipping point is only about 730,000 users. So if $730{,}000 < A(0) < 890{,}000$, applying full harm reduction can tip the model to a high-level equilibrium with 5.78 million users, rather than having use decay back down toward zero.

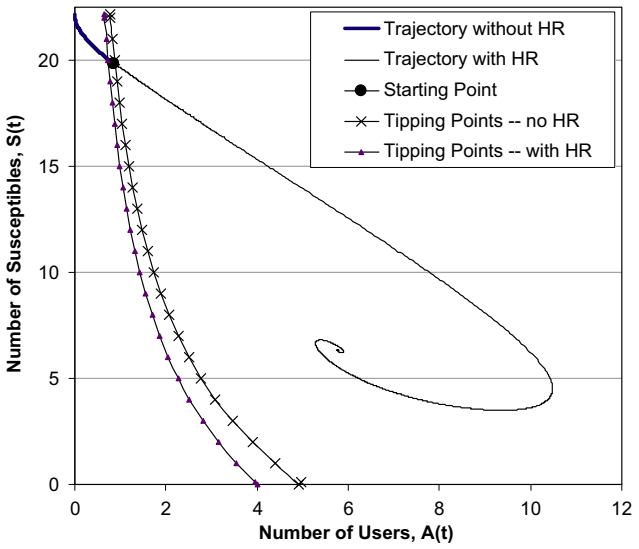

Fig. 2. Epidemic trajectories in the $A - S$ plane with and without harm reduction, starting at the same initial point (near vertical curves show tipping points separating regions of convergence to low- vs. high-levels of use, with and without harm reduction), numbers in million.

Note: the region where harm reduction increases total harm is broader than the region where it tips the epidemic toward the high-level steady state. For $A(0)$ in the two shoulder regions (600,000–730,000 & 890,000–1,200,000), implementing harm reduction does not "tip" the epidemic, but it does adversely affect the transient so severely that it increases total harm.

Fig. 2 illustrates the tipping point for the U.S. cocaine epidemic more clearly by contrasting the two trajectories emanating from a single initial condition ($A(0) = 850{,}000$, $S(0) = 19{,}843{,}000$) with harm reduction (the curl down to the lower right) and without harm reduction (the short segment moving up and to the left toward the vertical axis). Annotating the trajectories with time stamps makes the graph cluttered. Roughly speaking, however, the doubling time for the number of users on the increasing trajectory to the right of the tipping point is about 4–5 years. The "half life" of the number of the users when moving to the left of the tipping point is more sensitive to when one starts counting, but as an example the left-hand trajectory takes 15 years to decay from 0.6 to 0.3 million users.

That harm reduction – even the relatively modest sort available for US cocaine where v_{max} is only 17.4% – could so dramatically affect the trajectory of a drug epidemic might seem to be a sobering caution against using harm reduction. On the other hand, for this model implementing a harm reduction policy could adversely tip the epidemic only if done within a quite small region in $A - S$ space. Fig. 2 identifies that region by the two almost vertical curves, which are the stable paths leading to the saddle point equilibriums. The right hand of those two curves is the set of tipping points for the current model when there is no harm reduction. The left-hand curve shows how far that set of tipping points gets moved over when harm reduction is implemented as aggressively as possible ($v = v_{max} = 17.4\%$).

Fig. 2 has something for both sides of the harm reduction debate. Opponents of harm reduction can look at the two trajectories diverging from the common starting point and see the potential for substantial adverse effects on levels of use if harm reduction is fully applied starting from certain initial conditions. Proponents can look at how close the two sets of tipping points are and counter that the set of initial conditions for which that worst case scenario might come to pass is a quite narrow sliver in $A - S$ space.

Various sensitivity analyses can be performed, but dependence on v_{max} may be of particular interest. Results are more dramatic for the US case because it has two equilibria. Briefly, reducing v_{max} does not change much because v_{max} is already small. However, increasing v_{max}: (1) Moves the high-use steady state to the lower right (more users and fewer susceptible) because the decision maker can tolerate more users when they are less harmful; (2) Shifts the tipping point curve to the left because drug use is less harmful, so the current number of users has to be smaller in order for eradication to be optimal; and (3) Broadens the region where implementing harm reduction can tip the epidemic to the high steady state. The last effect is not so much because the system has become more sensitive, but rather because the "shove" of moving from use reduction to harm reduction now gives a bigger "push" to initiation and so can more easily "tip" the system over into the other basin of attraction.

4.2. Results for a hypothetical epidemic for which harm reduction is riskier

The previous subsection represented our best attempt to parameterize two real epidemics. However, as Appendix A makes clear, these parameter estimates are anything but precise. Here we pursue a sensitivity excursion by exploring whether changes to just a few parameters can make harm reduction a poor choice for a broader range of initial conditions. The parameter changes

are selected to make the sensitivity analysis dramatic, not by reference to any actual epidemic data.

In brief, the answer is yes. Relative to the base case US cocaine parameters in Table 1, all that would be required is for users to recognize fully the reductions in harm (ω is doubled from 0.5 to 1) and for the epidemic to be less virulent (i_2 declines from 0.009 to 0.00725). The latter change is of particular interest because of evidence that cocaine's "infectivity" has dropped in these latter parts of the US epidemic relative to its earlier value (Tragler et al., 2001; Caulkins et al., 2004).

With those changes full harm reduction $v = v_{max}$ increases total harm for any $A(0) > 550{,}000$ (about 12.5% of the new, lower no-HR steady state of 4.47M). So instead of being a suboptimal choice only within a certain, narrow range of initial conditions relatively early in the epidemic trajectory, harm reduction becomes counterproductive at any point much beyond those early stages. Even if the epidemic had stabilized at the no-HR high-level steady state, with these revised parameters implementing harm reduction still increases harm, albeit only slightly (use up by 24%, harm up by 3%). So it is not necessarily the case that harm reduction is always a good strategy once the epidemic has stabilized at endemic levels. Whether it is or is not depends on the particulars of the harm reduction policy, and is more likely to be appealing if users do not recognize fully all of the benefits that harm reduction brings them in terms of reduced cost of using drugs.

4.3. Implications of uncertainty in parameter values

The model greatly simplifies the epidemic dynamics, and in reality policy makers would never have perfect knowledge of model parameters. Hence, the policy makers' problem is actually quite a bit more complicated than is Formulation (7), involving stochastic optimization over parameter space. Fig. 3 hints at how such uncertainty could change the policy prescription, particularly for a risk-averse decision maker.

Fig. 3a contrasts the original tipping point curves with and without harm reduction (dashed and solid light lines) with the tipping point curves for a revised parameterization ($i_2 = 0.007$) with

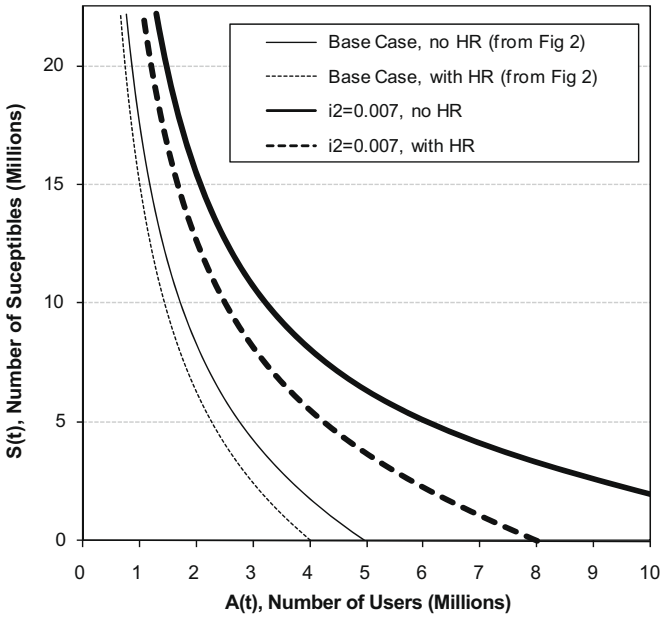

Fig. 3a. The tipping point curves both with (dotted lines) and without (solid lines) harm reduction move well to the right when the infectivity parameter i_2 declines modestly from 0.009 (light lines) to 0.007 (bold lines).

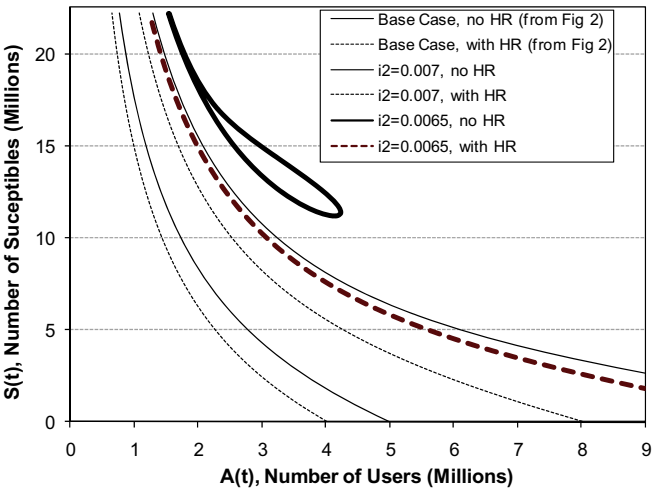

Fig. 3b. A modest further reduction in infectivity down to $i_2 = 0.0065$ dramatically alters the no-harm reduction tipping curve (bold solid line) so that the region in $A - S$ space with convergence to the high-level equilibrium shrinks to a small pocket (tipping curves from Fig. 3a shown as light lines for comparison).

and without harm reduction, drawn as dashed and solid heavy lines. Two things emerge.

First, the region in $A - S$ space where harm reduction could tip an epidemic toward the high-use equilibrium is still narrow, but is no longer just a sliver. Second and more importantly, that region has shifted substantially to the right in the $A - S$ plane. Naturally, for intermediate parameter values, the region where harm reduction could tip the epidemic is between the two regions shown in Fig. 3a. Hence, if one did not know the value of i_2 precisely and only knew that $0.007 < i_2 < 0.009$, then it would be possible for implementing harm reduction to tip the epidemic if the initial conditions were anywhere between the far left and far right curves.

Nor is the far right hand curve of Fig. 3a a firm bound on the region where harm reduction could potentially tip the epidemic. Fig. 3b contrasts the two pairs of tipping point curves already discussed (light lines) with two new ones obtained by decreasing i_2 just a bit further, from 0.007 down to 0.0065. That change pushes the tipping point curve with harm reduction only slightly to the right (dashed bold line), but it dramatically alters the tipping point curve when there is no harm reduction (solid bold line). No longer does the curve divide the relevant region of $A - S$ space in half, with initial conditions to the lower left approaching the no use equilibrium and those to the upper right approaching the high-use equilibrium. Rather, without harm reduction, the epidemic would approach the no use equilibrium unless the initial conditions happened to fall within a thin pocket surrounding the high-use equilibrium. Furthermore, decreasing i_2 a bit more makes even that thin pocket disappear.

The implications of Fig. 3 are clear, even without formally solving a stochastic optimization problem. There are regions in state space where harm reduction can dramatically increase future use. Since the harm reduction modeled in this subsection had $v_{max} = 17.41\%$, those dramatic increases in use also bring dramatic increases in total social cost. Furthermore, the location of those regions can be highly sensitive to certain parameter values. Hence, a risk-averse decision maker who wants absolute guarantees that a policy action will "do no harm" has some reason to shy away from an irrevocable implementation of harm reduction.

This cautionary tale by no means rules out harm reduction for all epidemics and all decision makers. If a decision maker is confident that a policy intervention will not tip the epidemic toward a much higher level of use, then harm reduction is less risky. Inasmuch as

use of many of the major drugs in many countries has more or less stabilized around endemic levels, decision makers might have confidence that the epidemic is not near a tipping point even if they do not have a fully parameterized model of the epidemic.

Likewise, if harm reduction could quickly be reversed when and if use increases more than expected, then there is less danger. That suggests that studying the optimal dynamic control version of Formulation (7) would be of interest.

Fig. 3b offers one more insight: implementing harm reduction can still tip an epidemic even when use is rapidly increasing before harm reduction is adopted. In a one-dimensional, continuous-time tipping point model, if use is increasing then the system must already be "to the right" of the tipping point and on a trajectory leading to the high-level equilibrium. That logic does not apply, however, to higher dimensional models. Consider, for example, what happens when one starts with 10 million susceptibles and 3 to 4 million users in Fig. 3b, with no harm reduction and $i_2 = 0.0065$. Those initial conditions are outside the "pocket" (solid bold line) delimiting initial conditions that lead to the high-level equilibrium, so we know the epidemic will eventually approach the lower equilibrium with no use. It does so by initially moving to the lower right (increasing use and decreasing numbers of susceptibles) and can go quite far to the right before curling back. Hence, it is not safe to presume that increasing use implies convergence to a high-level equilibrium and, hence, that there is no great risk from implementing harm reduction. It is possible that in such circumstances harm reduction could convert a "flash epidemic" with only transitory widespread use into one that stabilizes at high endemic levels.

5. Discussion

Drug abuse can do enormous damage to users and to society more generally, so it is natural to ask whether drug use can be made safer. To varying degrees the answer is, "Yes, drugs can be made safer, although by no means safe". Some countries (e.g., Australia) have embraced such "harm reduction" interventions. Others (e.g., the US) eschew harm reduction, in no small part because opponents worry that making drugs safer might encourage greater use.

Sensitive to this possibility, MacCoun (1998) reviews evidence concerning "behavioral compensation" in a variety of domains and suggests that people do often decide to participate in an activity more frequently when it is safer, but the increases are smaller, proportionately, than the reductions in harm, so total harm is generally reduced when an activity is made less harmful.

That the direct behavioral response is less than proportionate need not, however, imply that the total long-run change in use is less than proportionate if there is feedback in the system governing the dynamic evolution of use. It is generally accepted that social-interactions and/or market effects can generate such feedback in trajectories of drug prevalence. Hence, it is worth moving beyond a simple, static conceptualization that total harm = use * average harmfulness, to embed harm reduction interventions within a dynamic model of the evolution of drug use, as we have done here. Furthermore, as future research one may wish to explore a dynamically controlled version of this optimization model and/or richer versions of the model with more states and parameters to capture better the subtleties of population mixing and other refinements.

Nevertheless, the results here are a proof by example that in a system with feedback, reductions in harm that have a less than proportionate direct effect on initiation can still have a much more than proportionate effect on the present value of future use. It is noteworthy that the epidemic model employed was an exceedingly simple one. It has just two states (users and non-users who are susceptible to initiation) and only one non-linearity – namely that initiation arises from the mixing of users and susceptible non-users, with the probability that such interactions generate a new initiation possibly depending on whether the drug is widely or not so widely used. Hence, one does not have to concoct a particularly exotic set of non-linear feedbacks for dynamics to matter.

A particular mechanism that emerged from this simple model of drug use dynamics was a tipping point separating two stable equilibria, one with low-levels of use and the other with a much higher level of use. If the system is initially close to the tipping point but on the side leading to low-levels of use, implementing harm reduction might shift the tipping point, so that the current level of use is then on the side of the tipping point leading use to grow toward the high-level equilibrium.

This mechanism emerged in the model parameterized for US cocaine principally because the parameters suggest that random interactions between users and susceptible non-users were *more* likely to lead to initiation when cocaine use was common. In contrast, the parameterization for Australian IDU suggests that the likelihood such interactions lead to an initiation is *decreasing* in the current prevalence of use. This difference accounts in no small measure for the model result that harm reduction is always a good idea for Australian IDU, but there is a (relatively narrow) range of initial conditions for which harm reduction increases not just future use but also total harm for the model parameterized for US cocaine use. Those initial conditions involved levels of use far below current levels, so if the model were to be interpreted literally, it would suggest that implementing harm reduction for US cocaine today would reduce the present value of total future harm.

The models are highly stylized and parameterizations tenuous, so it is important not to put too much stock in those specific recommendations. However, many models of drug epidemics have tipping points and for various reasons (e.g., "enforcement swamping"), not just because of the specific mechanism in play here. So the caution about harm reduction is a more general one.

The model explored above also serves as a proof by example of the idea that the location of these tipping points can be highly sensitive to certain parameter values. Since in practice parameter values are not known with great precision or certainty, this suggests an additional source of caution.

Hence, the appropriate policy prescription becomes, "One should be more cautious about implementing harm reduction when the drug problem may be near a point where modest perturbations favoring greater use can be multiplied into large changes in use".

That is subtly but importantly different than what we expected. We expected the bottom line recommendation to be, "Harm reduction can safely be implemented late in an epidemic, when use has stabilized near endemic levels". However, we found – without great effort – examples where feedback can make harm reduction counter-productive even when use had already stabilized at endemic levels.

A more general implication is that the bitterly opposing sides of the harm reduction debate may both have valid points. Common ground – or at least a more productive characterization of differences – might come from both sides articulating more carefully their presumptions concerning the dynamics of drug use. Then both sides could couch their recommendations in conditional language, conditional on those presumptions. For example, a harm reduction advocate might say, "Harm reduction is worth considering for this drug in this country because of the following evidence concerning its dynamics" rather than saying "Harm reduction is unequivocally and universally the best policy".

A final point concerns the modeling strategy. Many product diffusion models have been explored. Notable among them are the models by Easingwood et al. (1983), who extend the original Bass

model allowing for non-uniform influence between buyers, A, and the remainder of the population, $1 - A$, that have not yet adopted. It is clear this perspective can generate interesting dynamics. These dynamics were explored here for controlling the spread of a "bad" consumption good, whose use is penalized in the objective function. One could easily imagine applying the same principle to marketing models where the objective is to encourage the spread of a good whose consumption is beneficial to the decision maker.

An analog of harm reduction in marketing would be raising prices, which improves the per person objective function value but over time can have an adverse effect on the population of users (customers). One could also explore the inverse problem of issuing coupons or price discounts that promote use, but diminish the instantaneous impact of that use on the objective function (by reducing net revenues on sales at the discounted price). The analogous managerial recommendations could include that price discounts might be most useful if the adoption dynamics are near some tipping point separating decay to a low-level of use from an equilibrium in which the product is in widespread use.

Hence, avenues for future work include not only constructing similar models for other drugs or more elaborate state spaces representing different intensities of use and exploring dynamic control formulations. They also include "inverse" applications in marketing where high levels of use are good, not bad, for the decision maker.

Acknowledgements

This work was funded in part by the Austrian Science Fund (FWF, P18527-G14), a Robert Wood Johnson Foundation Health Investigator Award, the Qatar Foundation, and a Grant from the Qatar National Research Program. The views expressed imply no endorsement by any of these funders.

Appendix A. Derivation of parameter values

A.1. Drug use trajectory parameters

US cocaine parameters were derived from Caulkins et al.'s (2004) epidemic data. Specifically, the number of cocaine users was taken as the sum of the number of "light" and "heavy" users ($A(t) = L(t) + H(t)$) in Caulkins et al.'s Figs. 3 and 4, and Initiation Series 3B was used for the initiation series, $I(t)$. (Analysis was replicated with $A(t)$ defined to be a weighted average of light and heavy users weighting by relatively proclivities to consume, but the results were similar.)

The exit rate $\mu = 0.16607216769$ gives the smallest sum of squared errors (SSE) between the $A(t)$ data and a modeled $A(t)$, call it $A_m(t)$, obtained from $A_m(t) = I(t-1) + (1-\mu)A_m(t-1)$ started with $A_m(1972) = A(1972)$.

A series $S(t)$ was created with $S(1972) = k/\delta$ and $S(t) = k + (1-\delta)S(t-1) - I(t-1)$ for $t > 1972$. The k, δ, and initiation function parameters were found by minimizing the SSE between the Caulkins et al. $I(t)$ series and the series $(i_1 + i_2 A(t)^{\alpha})S(t)$, subject to the constraint that all parameters be non-negative, with that constraint binding only for i_1, which was estimated to be 0. The parameters in Table 1 are for $A(t)$ being the Caulkins et al. (2004) $A(t)$. Similar results were achieved using $A_m(t)$.

Australian data were based on Caulkins et al. (2007). That analysis focused on occasional and frequent injection drug users, not heroin users per se, but there is considerable overlap, so $A(t)$ and $I(t)$ were taken to be the estimated prevalence and inflow to that population, respectively. Almost no one starts drug use by injecting, so susceptibles are those already using drugs by means other than injection (i.e., the sum of Caulkins et al.'s cannabis only users ($C(t)$) and those using more than cannabis but not by injection, ($M(t)$).

Inflow into the pool of susceptibles is taken to be the product of average initiation into drug use times the proportion who would at some point consider injecting (estimated to be 0.2232 million and 23.5%, respectively). Their outflow rate δ is the inverse of the weighted average of dwell times for all people who initiate any drug use. Caulkins et al. estimated 8.8% initiate directly into the M-state, which has a dwell time of 8.6 years. The other 91.2% initiate into the C state, which has an average dwell time of 7.5 years, with 37.2% of them then escalating into the M state. So the overall average dwell time is $(8.8\% * 8.6 + 91.2\% * (7.5 + 37.2\% * 8.6)) = 10.5$ years, yielding a $\delta = 0.0952$.

These parameters were used to create two time series of susceptibles starting with initial values in 1960 at the possible extreme values of $\widehat{S} = k/\delta$ and 0. To fit the parameters for $(i_1 + i_2 A(t)^{\alpha})S(t)$, to the $I(t)$ series, we regressed $\log(I(t)/S(t) - i_1)$ on $\log(A(t))$ for various values of i_1. It turned out that $i_1 = 0$ gave the best fit, and the values of i_2 and α were similar whether $S(t)$ was produced with $S(1960) = \widehat{S}$ and 0, so Table 1 contains the midpoint of those two estimates. Of all the parameters estimated, those for initiation into Australian IDU are the most suspect because of data limitations and the limited extant literature that models that population quantitatively.

The exit rate from $A(t)$ is the weighted average of Caulkins et al. (2007) estimated exit rate for occasional (0.15) and frequent (0.05) injectors, weighting by their relative proportions (175,000 and 100,000, respectively, according to Law et al. (2003)), or $\mu = 0.1136$.

A.2. Parameterizing the effect of harm reduction on initiation

Caulkins et al. (2002) estimate the social cost per gram of cocaine consumed to be $215.18 in 2001 dollars, which is $223.56 per pure gram in 2003 dollars. That estimate was worked up from Harwood et al. (1998), with some adaptations, and Harwood et al. suggest that 17.408% of cocaine related social costs in the US are attributable to medical consequences and premature death, so that is taken as the value of v_{max}. Also, from Caulkins et al. (2004) we have that the average price per pure gram of cocaine (powder) purchased in <2 gram units in 2003 was $106.54.

The parallel estimates for Australia are much harder to come by because less has been published about social costs of drugs in Australia, because the only break out of social costs by type of user is done by drug use and state of dependence, not by injection, and because Australia already has been implementing some harm reduction.

The starting point is Moore and Caulkins's (2006) estimates of the health-, crime-, and road accident related social costs of drug use for Australia broken down by substance and by whether or not the user is dependent. The A users represent injection drug users. In Moore's terms, that might be construed as including both dependent and non-dependent heroin users and a subset of dependent amphetamine and cocaine users. McKetin et al. (2005) suggest that proportion might be about 76% for amphetamine users. We found no similar estimate for cocaine users, but cocaine use is relatively rare in Australia and accounts for a quite modest share of all social costs, so in the absence of any better figure, we apply the 76% to cocaine as well.

Table A1 summarizes the relevant information in Moore and Caulkins (2006, Tables 16 and 17) (e.g., including 76% of Moore's estimate of total health-related costs associated with dependent amphetamine users and 76% of his estimate of the number of dependent amphetamine users). Many who inject amphetamine or cocaine also inject heroin, so to adjust for double counting we presume the total number of injectors is the number of heroin users plus half of the dependent amphetamine and cocaine

Table A1
Social costs associated with those presumed to be Australian IDUs (Millions of Australian Dollars).

	Opiates (dependent use)	Opiates (Non-dependent use)	Amphetamine (dependent use)	Cocaine (dependent use)	Total
Health-related costs	$2,976	$55	$358	$87	$3,389
Crime-related costs	$1,242	$18	$1,588	$77	$2,848
Road accident costs	$261	$46	$154		$461
Total costs – dependence	$4,479	$119	$2,101	$164	$6,699
# of users	41,401	107,898	55,480	5,320	179,699

injectors. McKetin et al. (2005, p. 11) report that 72% of dependent (meth)amphetamine users in their sample "had a history" of heroin use. "Had a history of" is not precisely defined but seems to be use at some point in the past, not just in the past-year, so the overlap in past-year use must be less. So we used a proportion of 50% for this overlap. Hence we estimate the total number of IDUs contributing to Moore's social cost estimate to be 41,401 + 107,898 + (1 − 50%)*(55,480 + 5,320) = 179,699. This is considerably less than the Law et al. (2003) estimate of 275,000 IDUs in Australia, but the Law et al. figure came from before the dramatic reduction in heroin availability (the so-called "heroin drought", cf., Weatherburn et al., 2003), so the 176,699 figure is sensible. It suggests a total current social cost per IDU of $6699M/179,699 = $37,277 with health-related social costs per IDU per year of $3389M/179,699 = $18,859. Moore excludes both the costs of control and labor productivity effects, so this gives a somewhat optimistic view of the proportion of social costs that harm reduction can address, but one that is consistent with the US cocaine estimates above.

Complicating matters is that Australia already does some harm reduction. As a first-order approximation, it might be fair to say that Australia has essentially eliminated HIV/AIDS costs (cf., Moore and Caulkins, 2006). Drummond (2004) credits those programs with having averted 25,000 HIV infections which would carry an average annual post-diagnosis cost of $14,000 per year, so recent Australian efforts might already have been saving about $350 million per year, or about $1950 per IDU. Note: The long-run savings from these averted HIV infections will be much larger, but these reflect more closely savings in the year of Moore's analysis.

Adding this additional $1950 per IDU in health-related costs raises the total health-specific and all-social costs per IDU per year in the absence of any harm reduction from $18,859 and $37,277 to $20,809 and $39,227, respectively. This suggests setting v_{max} = $20,809/$39,227 = 0.53 for Australian IDU.

To estimate c_m, the monetary cost of injection drug use to the user, we start with Moore (2005, p. 22) observation that the consensus in the literature was that "regular" heroin users spend $600 per week on heroin and that they consumed 17.5 times as much as did occasional users, suggesting average annual spending on drugs by IDU of about (41,401 + 50% * (55,480 + 5320) + 107,898/17.5) * $600 * 52/179,699 = $13,537.

Finally, concerning the elasticity of participation, Dave (2004, p. 26) reports that the "long-run price elasticity of the probability of participating in each drug [are] about −0.45 for cocaine and −0.21 for heroin" based on US data. There are no comparable studies of heroin elasticity in Australia, let alone for all Australian IDU, so we use the −0.21 figure for the Australian model as well.

Appendix B. Analysis of the stability of the equilibria

The Jacobian matrix for the systems dynamics is

$$B = \begin{pmatrix} -(\delta + f(A)h(v)) & -f'(A)h(v)S \\ f(A)h(v) & f'(A)h(v)S - \mu \end{pmatrix}$$

so

$$sI - B = \begin{pmatrix} s + (\delta + f(A)h(v)) & f'(A)h(v)S \\ -f(A)h(v) & s + \mu - f'(A)h(v)S \end{pmatrix}.$$

Setting the $det|sI - B|$ to zero gives the discriminant equation

$$s^2 + (\delta + \mu + f(A)h(v) - f'(A)h(v)S)s + \mu(\delta + f(A)h(v)) - \delta f'(A)h(v)S = 0.$$

B.1. Case A: steady state with $\widehat{A} = 0$ and $\widehat{S} = k/\delta$

If $g()$ is a power function, then $f'(A)|_{A=0}$ is infinite (zero) if $\alpha < 1 (\alpha > 1)$, implying that the coefficients of s^1, and s^0 are both negative (positive), respectively, implying that the steady state with $\widehat{A} = 0$ is either semi-stable or stable, respectively.

B.2. Case B: interior steady state with concave f()

From Eq. (7), in any steady state, $\mu\widehat{A} = f(\widehat{A})\widehat{S}h(v)$, so as long as $f(\widehat{A}) > 0$, we can rewrite the discriminant equation as

$$s^2 + \left(\delta + f(A)h(v) + \mu\left(1 - \frac{f'(A)A}{f(A)}\right)\right)s + \mu f(A)h(v) + \mu\delta\left(1 - \frac{f'(A)A}{f(A)}\right) = 0.$$

Since $f(A) > 0$, if $f()$ is concave so $\left(1 - \frac{f'(A)A}{f(A)}\right) > 0$, then all three coefficients (of s^2, s^1, and s^0) are positive, implying that the roots are negative or are complex conjugate with negative real parts, so any interior steady state with $f()$ concave is stable.

B.3. Case C: 2 interior steady states with convex f()

This is the case where Eq. (10) crosses Eq. (9) 2 times. As Eq. (10) is monotonically increasing in $\widehat{A}$, there is a steady state $\widehat{A}_L$ with a lower A value and one with a higher A value $\widehat{A}_H$ compared to the critical value $\widehat{A}_C = \left(\frac{(1-\alpha)\delta}{h(v)i_2}\right)^{1/\alpha}$, where (9) hits (10) in a tangent point. Further, we see that $\frac{f'(A)A}{f(A)} = \alpha$ and $f(\widehat{A}_L) < f(\widehat{A}_C) = \frac{(\alpha-1)\delta}{h(v)}$, so the coefficient of s^0 is $\mu f(\widehat{A}_H)h(v) + \mu\delta(1-\alpha) < \mu f(\widehat{A}_C)h(v) + \mu\delta(1-\alpha) = \mu\delta(\alpha-1) + \mu\delta(1-\alpha) = 0$. So, the third coefficient of the discriminant equation is negative, and the first is positive. The sequence of coefficients changes sign once, so there is one positive Eigenvalue. The other eigenvalue is negative, so the low steady state $\widehat{A}_L$ is a saddle point.

For the high steady state, $f(\widehat{A}_H) > f(\widehat{A}_C)$ holds, so the coefficient of $s^0 > 0$. The coefficient of s^1 is $\delta + f(\widehat{A}_H)h(v) + \mu(1-\alpha) > \delta + f(\widehat{A}_C)h(v) + \mu(1-\alpha) = \delta + \delta(\alpha-1) + \mu(1-\alpha) = \delta\alpha + \mu(1-\alpha)$. The sign of this expression depends on the actual parameter values, but for the sets of parameter values analyzed it is positive. So all coefficients are positive and the high steady state $\widehat{A}_H$ is stable.

References

Bass, F.M., 1969. A new product growth model for consumer durables. Management Science 15, 215–227.

Baveja, A., Batta, R., Caulkins, J.P., Karwan, M.H., 1993. Modeling the response of illicit drug markets to local enforcement. Socio-Economic Planning Sciences 27, 73–89.

Behrens, D.A., Caulkins, J.P., Tragler, G., Haunschmied, J.L., Feichtinger, G., 1999. A dynamical model of drug initiation: Implications for treatment and drug control. Mathematical Biosciences 159, 1–20.

Behrens, D.A., Caulkins, J.P., Tragler, G., Feichtinger, G., 2000. Optimal control of drug epidemics: Prevent and treat – but not at the same time. Management Science 46, 333–347.

Blower, S.M., Koelle, K., Mills, J., 2002. Health policy modeling: epidemic control, HIV vaccines and risky behavior. In: Kaplan, Brookmeyer (Eds.), Quantitative Evaluation of HIV Prevention Programs. Yale University Press, pp. 260–289.

Bundgaard-Neilsen, M., 1976. The international diffusion of new technology. Technological Forecasting and Social Change 8 (4), 365–370.

Caulkins, J.P., 1993. Local drug markets' Response to focused police enforcement. Operations Research 41, 848–863.

Caulkins, J.P., 2005. Models pertaining to how drug policy should vary over the course of an epidemic cycle. In: Lindgren, B., Grossman, M. (Eds.), Substance Use: Individual Behavior, Social Interactions, Markets, and Politics, Advances in Health Economics and Health Services Research, vol. 16. Elsevier, pp. 407–439.

Caulkins, J.P., Reuter, P., 1997. Setting goals for drug policy: Harm reduction or use reduction. Addiction 92, 1143–1150.

Caulkins, J.P., Pacula, R., Paddock, S., Chiesa, J., 2002. School-based drug prevention: What kind of drug use does it prevent? MR-1459-RWJ. RAND, Santa Monica, CA.

Caulkins, J.P., Behrens, D.A., Knoll, C., Tragler, G., Zuba, D., 2004. Markov chain modeling of initiation and demand: The case of the US cocaine epidemic. Health Care Management Science 7 (4), 319–329. doi:10.1007/s10729-004-7540-4.

Caulkins, J.P., Dietze, P., Ritter, A., 2007. Dynamic compartmental model of trends in Australian drug use. Healthcare Management Science 10 (2), 151–162. doi:10.1007/s10729-007-9012-0.

Collins, D.J., Lapsley, H.M., 2002. Counting the cost: Estimates of the social costs of drug abuse in 1998–99. Commonwealth of Australia, Canberra.

Coomber, R., 2006. Pusher Myths: Re-Situating the Drug Dealer. Free Association Books, London.

Dave, D., 2004. Illicit Drug Use Among Arrestees and Drug Prices. NBER Working Paper #10648. Cambridge, MA. (http://ideas.repec.org/p/nbr/nberwo/10648.html).

Drummond, M., 2004. Return on Investment in Needle and Syringe Exchange Programs in Australia. Commonwealth Department of Health and Ageing, Commonwealth of Australia, Canberra.

Easingwood, C.J., Mahajan, J., Muller, E., 1983. A nonuniform influence innovation diffusion model of new product acceptance. Marketing Science 2 (3), 273–295.

Ferrence, R., 2001. Diffusion theory and drug use. Addiction 96, 165–173.

Gable, R.S., 2004. Comparison of acute lethal toxicity of commonly abuse psychoactive substances. Addiction 99, 686–696.

Grossman, M., 2004. Individual Behaviors and Substance Abuse: The Role of Price. NBER Working Paper No. 10948. Cambridge, MA.

Harwood, H., Fountain, D., Livermore, G., 1998. The Economic Costs of Alcohol and Drug Abuse in the United States, 1992. National Institutes on Drug Abuse, Rockville, MD.

Kirby, K.N., Petry, N.M., 2004. Heroin and cocaine abusers have higher discount rates for delayed rewards than alcoholics or non-drug-using controls. Addiction 99, 461–471.

Kleiman, M.A.R., 1988. Crackdowns: The effects of intensive enforcement on retail heroin dealing. In: Chaiken, M.R. (Ed.), Street-Level Drug Enforcement: Examining the Issues. National Institute of Justice, Washington, DC.

Kleiman, M.A.R., 1993. Enforcement swamping: A positive-feedback mechanism in rates of illicit activity. Mathematical and Computer Modeling 17, 65–75.

Kleiman, M.A.R., 1999. "Economic cost" measurements, damage minimization and drug abuse control policy. Addiction 94, 634–638.

Law, M.G., Dore, G.J., Bath, N., Thompson, S., Crofts, N., Dolan, K., Giles, W., Gow, P., Kaldor, J., Loveday, S., Powell, E., Spencer, J., Wodak, A., 2003. Modelling hepatitis C virus incidence, prevalence and long-term sequelae in Australia, 2001. International Journal of Epidemiology 32, 717–724.

MacCoun, R.J., 1998. Toward a psychology of harm reduction. American Psychologist 53, 1199–1208.

McCaffrey, B.R., 1998. Legalization would be the Wrong Direction. Los Angeles Times, July 27.

McKetin, R., McLaren, J., Kelly, E., Hall, W., Hickman, M., 2005. Estimating the Number of Regular and Dependent Methamphetamine Users In Australia. Technical Report 230. NDARC, Sydney.

Moore, T.J., 2005. Monograph No 01: What is Australia's "Drug Budget"? The policy mix of illicit drug-related government spending in Australia. DPMP Monograph Series. Fitzroy: Turning Point Alcohol and Drug Centre.

Moore, T.J., Caulkins, J.P., 2006. How cost-of-illness studies can be made more useful for illicit drug policy analysis. Applied Health Economics and Health Policy 5, 75–85.

Musto, D., 1987. The American Disease, second ed. Yale University Press, New Haven: CT.

Nordt, C., Stohler, R., 2006. Incidence of heroin use in Zurich, Switzerland: A treatment case register analysis. The Lancet 367, 1830–1834.

Noymer, A., 2001. The transmission and persistence of 'Urban Legends': Sociological application of age-structured epidemic models. Journal of Mathematical Sociology 25, 299–323.

Reuter, P., 1999. Are calculations of the economic costs of drug abuse either possible or useful? Addiction 94, 635–637.

Reuter, P., Caulkins, J.P., 1995. Redefining the goals of drug policy: Report of a working group. American Journal of Public Health 85 (8), 1059–1063.

Rigter, H., 2006. What drug policies cost: Drug policy spending in the Netherlands in 2003. Addiction 101, 323–329.

Riley, K.J., 1997. Crack, Powder Cocaine, and Heroin: Drug Purchase and Use Patterns in Six U.S. cities. National Institute of Justice, Washington, DC.

Tragler, G., Caulkins, J.P., Feichtinger, G., 2001. Optimal dynamic allocation of treatment and enforcement in illicit drug control. Operations Research 49, 352–362.

United Nations Office on Drugs and Crime, 2005. World Drug Report. Oxford University Press.

Weatherburn, D., Jones, C., Freeman, K., Makkai, T., 2003. Supply control and harm reduction: Lessons from the Australian Heroin 'Drought'. Addiction 98, 83–91.

Winkler, D., Caulkins, J.P., Behrens, D., Tragler, G., 2004. Estimating the relative efficiency of various forms of prevention at different stages of a drug epidemic. Socio-Economic Planning Sciences 38, 43–56.