Addiction

FOR DEBATE

doi:10.1111/j.1360-0443.2011.03461.x

Basing drug scheduling decisions on scientific ranking of harmfulness: false promise from false premises

Jonathan P. Caulkins[1], Peter Reuter[2] & Carolyn Coulson[3]

Carnegie Mellon University Heinz College and Qatar Campus, Pittsburgh, PA, USA,[1] School of Public Policy and Department of Criminology, University of Maryland, College Park, MD, USA[2] and Carnegie Mellon University Heinz College, Pittsburgh, PA, USA[3]

ABSTRACT

In recent years a number of studies have attempted to rank drugs by a single measure of harmfulness as the basis for decisions about scheduling and classification. These efforts are fundamentally flawed, both conceptually and methodologically. The effort to provide a single measure masks the variety of non-comparable dimensions that are relevant, the fact that benefits are ignored for most, but not all, drugs and that the harms of a drug are not invariant to the policy regime chosen. Methodologically, the most prominent recent effort ignores drug interactions and mixes aggregate and individual harms inappropriately. Instead we suggest that multiple dimensions of harm need to be displayed to inform human judgments of what drugs should be scheduled. Harm is not usefully reducible to a single dimension, and even perfect rankings would not constitute a 'sufficient statistic' for determining scheduling decisions.

Keywords Addiction, decision making, drug policy, harm indicators, scheduling.

Correspondence to: Jonathan P. Caulkins, Carnegie Mellon University Heinz College and Qatar Campus, 5000 Forbes Avenue, Pittsburgh, PA 15237, USA.
E-mail: caulkins@andrew.cmu.edu
Submitted 24 December 2010; initial review completed 24 January 2011; final version accepted 29 March 2011

INTRODUCTION

The most fundamental policy decision societies make with respect to a psychoactive substance is whether to 'schedule' it (effectively a form of prohibition) or to make it available, subject only to various regulations (taxes, labeling and quality regulations, etc.). Scheduled substances must be placed in a particular category, a choice that can determine sanctions and/or conditions of availability for medical use. A voluminous literature attempts to document how badly these decisions have been made in the past, both in terms of failed actions (notoriously, US alcohol Prohibition) and failed decision processes, with moral panic [1], racial prejudice [2] and sheer ignorance [3] having fouled deliberations over the last century.

Given this history, it is not surprising that scholars argue for science- or evidence-based scheduling decisions [4,5]. An important theme is objectively and quantitatively assessing the harmfulness of various substances (e.g. [6,7]), with the implicit or explicit premise that more harmful substances should be banned and less harmful substances should not be (cf. [8]).

That enterprise is, however, misguided in principle and in the particulars. Particular flaws of ranking systems are important to note and correct, but even a hypothetical perfect set of harm rankings would not constitute a 'sufficient statistic' for determining scheduling decisions. We lay the groundwork for discussing limitations of even perfect rankings by first distinguishing individual versus aggregate harm, and then commenting on particulars in the subsequent two sections.

INDIVIDUAL- AND AGGREGATE-LEVEL PERSPECTIVES SHOULD NOT BE MINGLED

Harmfulness to an individual and aggregate harm to society are entirely different concepts. Mingling the two is a common source of faulty reasoning because harm rankings can be entirely different at the individual versus aggregate unit of analysis. Being hit on the head by a meteorite is highly harmful to the individual, but meteorite strikes are a negligible source of population-level mortality. Similarly, taking cyanide is more harmful to an

individual than is taking ethanol, but ethanol abuse is a much larger problem for society.

The simplest way to distinguish macro- and micro-level harm is to think of the former as the latter multiplied by the extent of use [9]. Amount of use could have various meanings: prevalence or quantity (weight) consumed or even more refined measures (e.g. binges). The distinctions are discussed in MacCoun & Reuter ([10], p. 317–19) and need not concern us here.

The key point is that when national policy is made on utilitarian grounds, the focus is on aggregate outcomes, such as social welfare or total harm [11]. In contrast, the medical, laboratory or individual-user perspective focuses on harm per unit of use. We will call that 'harmfulness' to distinguish it from total or aggregate harm. Crucially, aggregate harm is not simply harmfulness scaled-up to the societal level, because amount of use is not a constant across substances, policies or circumstances. Indeed, inasmuch as harmfulness affects the extent of use, the relationship between harmfulness and aggregate harm is necessarily non-linear and, at least in theory, can even be inverse. Increasing harmfulness increases harm for individuals who do not moderate their consumption, but aggregate harm may or may not go up; that depends upon how much use is deterred by the increase in harmfulness. The interplay is complicated by externalities (harms to others) and benefits to users. Externalities influence aggregate harm but not the incentives of users; the case for prohibiting tobacco is weakened by the fact that such a large proportion of its harms fall on the users themselves. Alcohol causes so much societal harm in part because it offers so many benefits to its users.

A key point is that extent of use is also affected by policy—specifically, the decision to ban or allow a substance. Indeed, the primary purpose of prohibiting a substance is to reduce its use. Therefore, it is wrong to argue that legal substance A causes more total harm than does illegal substance B; ergo, B should be legal and/or A should be illegal.

Because decisions about national policies should be informed by aggregate outcomes, policy analysis has to consider the number of users throughout. One would think this is such an obvious point that it is not worth belaboring. Nevertheless, the most recent and perhaps most prominent article in the harm-ranking literature fails on precisely this point.

Nutt *et al.* [12] considered 16 harm criteria divided into two groups: nine relate to harms to users and seven to harm to others. They report and discuss harms on these 16 criteria and the part scores, but for present purposes our concern is their 'overall harm score', which is a weighted sum: $0.46 \times$ harm to users $+ 0.54 \times$ harm to others.

Nutt *et al.* [12] measure the first group on a per-person basis, and the second (mainly) in terms of aggregate harm. Thus Nutt *et al.*'s [12] overall harm scores measure neither harm per unit of use nor aggregate harm, but an awkward weighted average of each. The portion of the overall harm score reflecting harm per unit of use ignores harmfulness to others. Similarly, the portion reflecting aggregate harm ignores harm to users themselves. Thus, when these harm scores are brought into a policy debate, they enforce an implicit assumption that policy can have no effect on harm to users by changing the amount of use. Hiding total harm's dependence on prevalence in this way is analytical malpractice.

The mischief caused by this adding of apples and oranges is also revealed by comparing Nutt *et al.*'s scores for tobacco (26) and γ-hydroxybutyric acid (GHB) (19). They claim to have a ratio scale, so 26 can be thought of as being about one-third larger than 19. However, few people would believe that tobacco, with 8.5 million smokers in the United Kingdom [13] causes only one-third more harm to society than does GHB, used by only about 50 000 in the United Kingdom [14]. The explanation is that Nutt *et al.* [12] judge tobacco and GHB to be similarly harmful per user (37 each), and neither to be very harmful to others (17 versus 2); so tobacco's vastly higher prevalence is largely ignored because it is reflected only in the harm to others term.

EVEN PERFECT HARM RATINGS WOULD NOT BE ENOUGH TO CREATE POLICY

Suppose one somehow obtained perfect ratings of all drugs' aggregate harm to society, including both harms to users and to others: would such ratings constitute 'sufficient statistics' for an evidence-based process to determine whether and how each substance should be scheduled?

Superficially, one might think the answer is yes. One could rank the substances from greatest to least harm and prohibit those whose harms exceeded some threshold. Prohibited substances could then be placed in various categories based on rank. The thresholds might be determined by the democratic political process, whereas scientists could provide objective harm ratings, yielding a neat division of labor in the policy process.

That tidy vision is misguided for several reasons. First, the unit of analysis in policy modeling is the decision, not the substance. To simplify, the relevant comparisons are, for example, cocaine being scheduled versus cocaine not being scheduled; they are not cocaine versus amphetamine. Users may choose between drugs, but societies choose between policies.

Thus, scheduling decisions should be informed by the projected *change* in harm caused by the scheduling

change, not just the level of harm under a single policy context (e.g. the *status quo*). Nutt *et al.* [12] find alcohol to be the most harmful substance, so perhaps 'alcohol' would be the answer if the policy question were: 'A genie just granted the ability to magically make one substance's harms disappear. At which substance should we point the genie's wand?'. However, changing scheduling status does not make all harm disappear, and it can create new ones. People who have read and understood Nutt *et al.*'s [12] analysis can still oppose prohibiting alcohol without being obtuse or unscientific.

Indeed, conflating choice with object makes the entire assessment exercise ill-defined. Important harms include drug-related crime, environmental damage and the cost of police and prisons, yet none of those are characteristics of a chemical; they depend as well on legal status and programs implementing laws. Methamphetamines create environmental problems when produced in small, technologically primitive laboratories; this would disappear if they were legal.

A second reason we should not simply rank substances by harm and prohibit those exceeding a threshold is that scheduling decisions are interrelated; they cannot be made one drug at a time. For example, with Nutt *et al.*'s [12] overall harm scores, any threshold between 27 and 54 would imply banning crack but allowing cocaine. However, it is easy to make crack from powder cocaine, so allowing powder cocaine *de facto* implies high availability of crack.

More generally, policy should account for drug interactions. If mephedrone is predictably consumed with alcohol, then the assessed harms of mephedrone should reflect that use pattern, not the essentially clinical exercise of judging the effects of mephedrone alone.

Interactions occur at the market level, not just the individual level. Changing availability of one substance may also affect—positively or negatively—demand for other substances. Schedule status affects both dollar price [15] and non-dollar costs of using (e.g. time to find the drug), and there is a growing literature on cross-price elasticity of demand documenting how changes in the price of one drug can affect consumption of another (Jofre-Bonet & Petry [16] provide a review).

Thirdly, consequences of scheduling depend upon context. Alcohol prohibition in the United States in the 1920s created enormous problems with gangland violence; alcohol prohibition in contemporary Saudi Arabia does not. The amount of violence unleashed by criminal pursuit of cocaine market profits is greater in countries with higher availability of firearms (notably the United States) than in others. Prohibiting heroin in a country with a strong public health infrastructure and commitment to syringe exchange may have very different effects on human immunodeficiency virus/acquired immune deficiency syndrome (HIV/AIDS) than would a similar prohibition elsewhere.

An essential part of context is current prevalence. Enforcement swamping (i.e. increases in market size reducing the individual's risk of arrests [17]) and other non-linear dynamics imply that black markets have multiple stable equilibria [15]. Hence, illegal use can stabilize at low or high levels.

All other things equal—including degree of addictiveness and individual level harmfulness—it is much easier to maintain a prohibition against a rarely used substance than to impose a new prohibition on a substance that is already widely used. Enforcing prohibition of phencyclidine (PCP), GHB, and lysergic acid diethylamide (LSD) is not overly onerous for the simple reason that there are relatively few sellers or users of those drugs. In contrast, the tens of millions of existing alcohol users would make establishing a prohibition on alcohol tremendously costly, both directly and in terms of black market harms. A similar principle applies in reverse. One would expect legalizing marijuana to have less effect on use than legalizing cocaine, because marijuana is already at a high-volume equilibrium with widespread availability and low prices, relative to other sources of an hour of intoxication.

It is also worth noting that drugs' rankings are not the same with respect to every policy decision. If the goal is to reduce violence in Mexico, legalizing cocaine in the United States would do more than would legalizing marijuana, but the drugs' rankings would be reversed for decriminalization. Decriminalizing marijuana might erode Mexican market share by promoting home cultivation, but decriminalizing cocaine would exacerbate violence in Mexico by increasing US demand while leaving the supply network wholly illegal [18].

To be clear, there is nothing conceptually wrong with creating univariate measures of harmfulness to inform an individual's decisions as to what drug to try. Similarly, we take it as a given that scientific evidence should inform policy making, perhaps including univariate measures of aggregate harm by substance, e.g. by breaking down cost-of-illness estimates by substance. What is not reasonable is presuming that absence of a direct relationship between such measures and stringency of prohibition is prima facie evidence of flawed policy; no scalar measure of harm is a sufficient statistic for policy design.

IS IT EVEN POSSIBLE TO CREATE HARM RANKINGS?

The previous section argued that drug-by-drug harm ratings provide at best very incomplete information upon which to base scheduling decisions, even if those ratings were somehow perfect. Here we challenge the idea that it

is possible to create a perfect, objective, evidence-based composite measure of a drug's harm.

The essential challenge is that harmfulness is not naturally unidimensional. One drug may be more likely to promote violence, another more likely to addict (itself a distinct harm), a third more likely to produce fatal overdose, so it is not clear how to collapse a vector of ratings on different harm components down into a single number. For two reasons, we see this as insoluble, even with further research.

First, however scientific the assessment of each drug's harmfulness on each harm dimension, the weight or importance attached to those dimensions when compressing a vector of numbers down into a single scalar is inevitably a matter of judgment [11], so computing composite ranks is not a value-free exercise the public can delegate to scientific experts without concern about whether the scientists' values are representative of the electorate's. For example, Nutt *et al.*'s [12] committee judged that drug-specific mortality should be weighted 80% as heavily as drug-related mortality. That is not a statement about numbers of deaths; it is a statement about valuation, as the associated Advisory Council on the Misuse of Drugs (ACMD) document ([19], p. 10) makes clear: 'MCDA does not directly compare different kinds of harms, it compares the *preference values* associated with the harms . . . harm expresses a level of damage. Value, on the other hand, indicates how much that level of damage matters'.

One particular issue of values concerns whether, which and how to account for a drug's benefits. Alcohol indeed causes great harms; that is widely acknowledged. However, alcohol is now legal because weight is also given to its pleasures. The users' perceived benefits of other drugs are ignored in policy decisions. This is a statement of political realities; it should not be an assumption of an analytical exercise.

Secondly, even if everyone shared the same values, the validity of a linear additive summation across harm components depends on a certain type of independence across attributes called 'mutual preferential independence' [20]. That independence does not always hold, notably between dependence liability and any attribute reflecting harm to the user. A substance that induces dependence but has no ill effects (100 on dependence, 0 on the other attributes) is of no particular concern; it would be like water. A substance that is very dangerous but has no appeal is similarly of no particular concern; it would be like arsenic. However, a substance that both harms and engenders use despite those harms can be highly problematic. Dependence is an important component of a drug's ability to attract use despite adverse consequences. In short, there are non-linear interactions among the attributes, so their weighted sum is not a reliable measure of overall threat.

This last criticism pertains only to simpler models of value functions, such as the linear and additive form used by Nutt *et al.* [12]. There are more complicated models [20], although they are more difficult to elicit and interpret, and so may or may not be practical within the policy process.

CONCLUSION: STRESS HARM MATRICES, NOT HARM RANKINGS

This comment is unambiguously sour on the enterprise of creating unidimensional drug harm ratings. For technical reasons we find their pursuit quixotic. More fundamentally, even if perfect ratings could be created, they would not provide sufficient basis for making the policy choices motivating their creation. It simply does not follow logically that we should ban the substances with the highest harm scores, and allow the rest.

We wish, however, to be constructive, not merely nihilistic, and therefore suggest an alternative. It is certainly not a silver bullet, but perhaps could usefully supplement the usual armamentarium of policy analysis. The basic challenge is the multi-dimensionality of drug harms. Our suggestion is to embrace rather than suppress that multi-dimensionality because substances' relative harmfulness depends fundamentally on the (inevitably subjective and context specific) weights one places on different criteria.

Scientists could invent a scale and announce to policy makers that heroin is more harmful than tobacco; or they could invent a different scale and announce that tobacco is more harmful. However, policy makers would be better informed if scientists said that heroin kills many people directly via overdose, but even chronic use does minimal organ damage, whereas few people die of nicotine overdose even though smoking kills hundreds of thousands over time through lung cancer and heart disease.

Hence, we advocate creating harm matrices rather than harm ratings. Imagine a row for each substance and a column for each dimension of harm, so each substance is associated with a row of harm-type-specific ratings, not just a single number. The perennial prominence of summary measures such as *US News & World Report* rankings of American colleges and universities suggests an unshakable enthusiasm for composite measures, but it is the burden of thought leaders to resist such popular trends. Simple is good; simplistic is not.

Actually, we suggest going a step farther. As argued above, harms are not attributes of the chemical compound only, but also of the larger context. Hence, rows should not be identified with substances, but rather with substances in a particular context. To inform scheduling decisions, one needs at a minimum two rows per substance, one for the substance under *status quo* conditions

and another projecting harms for the same substance in the same jurisdiction but with a different scheduling status.

Even the harm matrices are not sufficient for guiding a scheduling decision. As we noted, values matter and there are interaction effects; changing the schedule status of one drug can affect harms associated with other drugs. However, that is not a problem as long as one views harm matrices as tools for educating policy makers, not as algorithms that can replace them and their judgment. When there is no pretence of a one-for-one link between an analytical exercise and a related policy decision, one is free to use tools more flexibly. For example, we think it would be instructive to show policy makers harm matrices contrasting a row for street heroin with a row for heroin prescribed via heroin maintenance, or a row for cocaine selling at $150 per gram with a row for cocaine in an otherwise identical context but selling at $50 per gram, even if those contrasts do not correspond to available policy choices.

Harm matrices in particular may or may not turn out to be useful, but the motivation inspiring them is broadly applicable. Scholars ultimately may have a more constructive impact on policy when they seek parsimonious devices for helping busy policy makers enrich their understanding of a complex issue, empowering those policy makers to pursue their ends more effectively than when scholars create simplistic pseudo-scientific scores that tilt towards particular policies, even if those policies appeal to the community of academic experts, or at least a subset thereof.

Declarations of interest

This work received no outside funding, and none of the authors are connected to the tobacco, alcohol, pharmaceutical or gaming industries or any organization substantially funded by them.

References

1. Reinarman C. The social construction of drug scares. In: Adler P. A., Adler P., editors. *Constructions of Deviance: Social Power, Context, and Interaction*. Belmont, CA: Wadsworth; 2006, p. 155–65. Published 1994.
2. Courtwright D. T. *Dark Paradise: Opiate Addiction in America before 1940*. Cambridge: Harvard University Press; 2001.
3. New Zealand Press Association. *National MP Falls Victim to water hoax. Stuff.co.nz*. 2007. Available at: http://www.stuff.co.nz/national/politics/38005 (accessed 30 August 2010; archived by Webcite at: http://www.webcitation.org/5v7ofvjnD (accessed 20 December 2010).
4. Randerson J. *Take Decisions on Drug Classification Out of Politicians' Hands, Say Advisers. Guardian*. 2008; Available at: http://www.guardian.co.uk/science/2008/nov/25/illegal-drugs-classification (accessed 22 July 2010; archived by Webcite at: http://www.webcitation.org/5jGjFn9Dh (accessed 24 August 2009).
5. Wood E. *et al*. *The Vienna Declaration*. 2010. Available at: http://www.viennadeclaration.com/the-declaration/ (accessed 31 January 2011; archived by Webcite at http://www.webcitation.org/5yYeZgX4w).
6. Nutt D. J., King L. A., Saulsbury W., Blakemore C. Developing a rational scale for assessing the risks of drugs of potential misuse. *Lancet* 2007; **369**: 1047–53.
7. Gable R. S. Comparison of acute lethal toxicity of commonly abused psychoactive substances. *Addiction* 2004; **99**: 686–96.
8. Room R. The dangerousness of drugs. *Addiction* 2006; **101**: 166–8.
9. MacCoun R. J. *The Psychology of Harm Reduction: Comparing Alternative Strategies for Modifying High-Risk Behavior*, vol. 6. Oakland, CA: Wellness Lecture Series; 1996.
10. MacCoun R. J., Reuter P. *Drug War Heresies: Learning from Vices, Times and Places*. New York: Cambridge University Press; 2001.
11. Kalant H. Drug classification: science, politics, both or neither? *Addiction* 2010; **105**: 1146–9.
12. Nutt D. J., King L. A., Phillips L. D. Drug harms in the UK: a multicriteria decision analysis. *Lancet* 2010; **376**: 1558–65.
13. Department of Health. *Tobacco. UK: The National Archives*. Updated 9 April 2010. Available at: http://webarchive.nationalarchives.gov.uk/+/www.dh.gov.uk/en/Publichealth/Healthimprovement/Tobacco/index.htm (accessed 20 November 2010; archived by Webcite at: http://www.webcitation.org/5v7ocGGy3 (accessed 20 December 2010).
14. Hoare J., Moon D. *Drug Misuse Declared: Findings from the 2009/10 British Crime Survey: England and Wales*. London: Home Office; 2010, Available at: http://rds.homeoffice.gov.uk/rds/pdfs10/hosb1310.pdf (accessed 20 November 2010; archived by Webcite at: http://www.webcitation.org/5v7oYJ8JH (accessed 20 December 2010).
15. Caulkins J. P., Reuter P. How drug enforcement affects drug prices. In: Tonry M., editor. *Crime and Justice—A Review of Research*, vol. 39. Chicago, IL: University of Chicago; 2010, p. 213–72.
16. Jofre-Bonet M., Petry N. M. Trading apples for oranges? Results of an experiment on the effects of heroin and cocaine price changes on addicts' polydrug use. *J Econ Behav Organ* 2008; **66**: 281–311.
17. Kleiman M. A. R. Enforcement swamping: a positive-feedback mechanism in rates of illicit activity. *Math Comput Model* 1993; **17**: 65–75.
18. Kilmer B., Caulkins J. P., Bond B. M., Reuter P. *Reducing Drug Trafficking Revenues and Violence in Mexico: Would Legalizing Marijuana in California Help?* Santa Monica, CA: RAND; 2010, OP-325-RC.
19. Advisory Council on the Misuse of Drugs. *Consideration of the Use of Multi-Criteria Decision Analysis in Drug Harm Decision Making*. London: Home Office; 2010, Available at: http://www.homeoffice.gov.uk/publications/drugs/acmd1 (accessed 16 November 2010; archived by Webcite at: http://www.webcitation.org/5v7oUK2k0).
20. Keeney R. L., Raiffa H. *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*. New York: Wiley and Sons; 1976.

