'Dark logic': theorising the harmful consequences of public health interventions

Chris Bonell,[1] Farah Jamal,[1] G J Melendez-Torres,[2] Steven Cummins[3]

▸ Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/jech-2014-204671).

[1]Department of Childhood, Families and Health, Institute of Education, University of London, London, UK
[2]Department of Social Policy and Intervention, Centre for Evidence-Based Intervention, University of Oxford, Oxford, UK
[3]Department of Social & Environmental Health Research, London School of Hygiene & Tropical Medicine, London, UK

Correspondence to
Professor Chris Bonell, Department of Childhood, Families and Health, Institute of Education, University of London, 18 Woburn Square, London WC1H 0NR, UK; c.bonell@ioe.ac.uk

Received 15 July 2014
Revised 18 September 2014
Accepted 2 November 2014
Published Online First 17 November 2014

To cite: Bonell C, Jamal F, Melendez-Torres GJ, *et al*. *J Epidemiol Community Health* 2015;**69**:95–98.

ABSTRACT

Although it might be assumed that most public health programmes involving social or behavioural rather than clinical interventions are unlikely to be iatrogenic, it is well established that they can sometimes cause serious harms. However, the assessment of adverse effects remains a neglected topic in evaluations of public health interventions. In this paper, we first argue for the importance of evaluations of public health interventions not only aiming to examine potential harms but also the mechanisms that might underlie these harms so that they might be avoided in the future. Second, we examine empirically whether protocols for the evaluation of public health interventions do examine harmful outcomes and underlying mechanisms and, if so, how. Third, we suggest a new process by which evaluators might develop 'dark logic models' to guide the evaluation of potential harms and underlying mechanisms, which includes: theorisation of agency-structure interactions; building comparative understanding across similar interventions via reciprocal and refutational translation; and consultation with local actors to identify how mechanisms might be derailed, leading to harmful consequences. We refer to the evaluation of a youth work intervention which unexpectedly appeared to increase the rate of teenage pregnancy it was aiming to reduce, and apply our proposed process retrospectively to see how this might have strengthened the evaluation. We conclude that the theorisation of dark logic models is critical to prevent replication of harms. It is not intended to replace but rather to inform empirical evaluation.

INTRODUCTION

'First do no harm' is an ethical imperative above even doing good.[1] Although we might assume that public health programmes involving sociobehavioural rather than clinical interventions are unlikely to generate iatrogenic effects, it is well established that they sometimes cause serious harms.[2] Public health interventions involve human agency and are interruptions to complex social systems, so it is unsurprising that unintended effects can occur.[3] Popper stressed the importance of 'social engineering' being piecemeal and subject to empirical analysis regarding intended and (especially) unintended consequences.[4] Merton and Giddens have also drawn attention to the significance of unintended consequences of social programmes, often using the concept as an analytical tool through which to identify sociological forces at work. For these theorists, tracing unintended social repercussions is perhaps the most crucial element in the study of social phenomena.[5 6] However, the assessment of unintended and adverse effects remains a neglected topic in evaluations of public health interventions[7] other than in areas such as suicide prevention and illicit drug interventions.[8 9] Since harms are generally not measured in a consistent manner across studies, they are rarely examined in systematic reviews.[7 10] This is problematic as some harms are insufficiently common to be detected by single studies, but could be detected by meta-analyses.

More recently, interest in the potential harmful effects of public health intervention has increased, with attempts made to categorise types of harm.[2 7] Lorenc and Oliver offer the following typology: direct harms (eg, sports participation causing injuries); psychological harms (eg, screening producing stressful false-positive results); equity harms (eg, health promotion most benefiting those with the least need); group and social harms (eg, targeted interventions reinforcing risk by labelling or aggregating at-risk individuals); and opportunity harms (eg, ineffective interventions taking resources from more effective ones). In this paper, we use 'harm' more narrowly to mean harms which differentially affect individuals receiving an intervention and are common enough to detect in evaluations or meta-analyses, excluding very rare side effects (because even syntheses are unlikely to establish whether these are caused by interventions), opportunity harms (because these do not directly harm recipients) and inequities in intervention benefits (because these may arise from all benefiting, albeit differentially).[11 12] Our definition includes what pharmacologists term 'paradoxical effects', that is, interventions increasing adverse outcomes they seek to prevent,[13] and 'harmful externalities', where interventions produce harms in other outcomes.

In this paper, we first argue for the importance of evaluations of public health interventions not only aiming to detect potential harms but also the mechanisms that might underlie these harms so that they might be avoided in future. In doing so, we refer to the evaluation led by one of us (CB) of a youth-work intervention which unexpectedly appeared to increase the rate of teenage pregnancy it was aiming to reduce. Second, we examine empirically whether protocols for the evaluation of public health interventions do examine harmful outcomes and underlying mechanisms and, if so, how. Third, we suggest a new process by which evaluators might develop 'dark logic' models to guide the evaluation of potential harms and underlying mechanisms, and apply this retrospectively to the evaluation of the youth-work intervention to see how this might have strengthened the evaluation.

J Epidemiol Community Health: first published as 10.1136/jech-2014-204671 on 17 November 2014. Downloaded from http://jech.bmj.com/ on January 19, 2022 by guest. Protected by copyright.

RATIONALE FOR ASSESSING MECHANISMS UNDERLYING HARMS

One motivation for this paper was an evaluation one of us (CB) led of the Young People's Development Programme (YPDP). This example illustrates the importance of evaluations exploring not only potential harmful outcomes but also underlying mechanisms. Informed by effective youth-development programmes from the USA,[14] YPDP was delivered across England with the aim of reducing teenage pregnancies, drug use and school exclusions. Young people aged 13–15 years, whom teachers, social workers and other professionals identified as at risk of these adverse outcomes, were referred to their local programme. Each local site employed youth workers to provide recipients with additional education, arts and sports activities, mentoring and other components. A cluster randomised controlled trial (RCT) was impossible because intervention sites had been selected by competitive tender before evaluation started. An individual RCT was impossible because programme developers felt that allocating individuals to intervention/control groups would disrupt existing friendship groups. The evaluation was therefore quasi-experimental, prospectively comparing 27 YPDP sites with 27 control sites matched by evaluators on region, deprivation and teenage pregnancy rates. Young people in control sites (n=1087) were recruited using similar processes and criteria as YPDP recruitment (n=1637), and were followed up at 9 and 18 months to examine self-reported outcomes.

Much to the evaluators' surprise, even adjusting for multiple prehypothesised confounders, there were nearly four times as many pregnancies among girls in the intervention group than in the control group,[15] almost three times as many young people engaging in sex and over twice as many young people truanting from school. Attrition in the study was high because of the challenges in following up very vulnerable young people, but weighting increased the ORs of adverse outcomes. While recognising the problems arising from the non-randomised design and attrition, evaluators concluded that the intervention was at best ineffective and probably harmful, because of the large effect sizes that remained in all analyses. The evaluators developed post hoc ideas about *how* the intervention might have caused harm, but had developed no a priori hypotheses about these and thus could not examine them quantitatively.

Other evaluations have similarly reported harms, with varying degrees of clarity about the underlying mechanisms.[2 7] A group intervention for men who have sex with men aimed to develop attitudes and norms supportive of sexual risk reduction, but was instead associated with an increased risk of sexually transmitted infections. While this might have arisen because the intervention modified sexual risk networks, this was not examined empirically.[16] In contrast, evaluations of some drug-prevention interventions have noted adverse effects on the use of drugs and provided evidence that these may be mediated by interventions bringing recipients into contact with more risk-involved peers and reinforcing pro-risk attitudes and behaviours.[17] All the evaluations cited above detected 'paradoxical effects' rather than 'harmful externalities'. It is quite possible that because evaluators rarely aim to develop a priori ideas about the broader potential harmful effects of interventions, such harmful externalities may be inadequately detected in evaluations.

A better understanding of the underlying mechanisms could help ensure future interventions avoid iatrogenic mechanisms. For example, there is evidence that some youth group interventions which aim to reduce behaviours such as drug use can exacerbate risks, and that this occurs because the interventions are insufficiently structured, thus allowing positive peer reinforcement of pro-risk attitudes and behaviours.[2] This evidence of the underlying mechanisms is useful in ensuring that other youth interventions are better structured, thus avoiding positive reinforcement. Establishing that a particular intervention can cause harm does not necessarily mean that the theory of change or means of delivery of the intervention is wholly abandoned. A deeper understanding of the mechanisms underlying harms might enable further refinements of the sort Popper envisaged with his idea of piecemeal social engineering.[4] The next section explores whether protocols of public health evaluations aim to explore the potential harms and underlying mechanisms and, if so, using what approaches.

REVIEWING EXISTING PROTOCOLS

On 14 April 2014, we checked all projects funded since 2010 by the National institute of Health Research Public Health Research programme,[18] the major UK funder of public health evaluations. For all primary evaluations, experimental and quasi-experimental (including natural experiments), for which a protocol was provided, we reviewed what design the study employed and whether the protocol made any reference to potential harms. We assessed whether there was provision for 'harmful externalities' or merely 'paradoxical effects' to be evaluated (eg, using additional quantitative measures or qualitative research). We also reviewed whether the study aimed to examine pathways underlying potential harms (using quantitative mediator/moderator relationships or qualitative research). Results are reported in online supplementary appendix 1.

There were 29 protocols for trials, 12 for quasi-experimental (non-random controlled before/after studies), one before/after study and one laboratory study. Fourteen studies did not mention harm at all. Seventeen studies mentioned harm but aimed to examine this only in terms of paradoxical effects on primary or secondary outcomes. Nine of the 27 RCTs, 2 of the 11 quasi-experimental studies and 1 before/after study aimed to examine other harms, 10 using additional quantitative measures and 6 using qualitative research. Only one study aimed to examine the mechanisms underlying harms, using qualitative research.

It seems clear then that currently, at least in the UK, evaluations of public health interventions are inconsistent in their focus on potential harms and very few are focused on exploring the underlying mechanisms. Non-experimental studies appear to be particularly lacking in consideration of harms. The lack of attention to harms overall may be because evaluators lack a framework for hypothesising harms and associated mechanisms, whereas they do possess theories of change and logic models to guide their evaluation of intended intervention mechanisms and outcomes. The next section proposes 'dark logic' models as a systematic process for prehypothesising what harms and underlying mechanisms might plausibly arise for particular interventions.

DEVELOPING 'DARK LOGIC' MODELS OF POTENTIAL INTERVENTION HARMS AND THEIR UNDERLYING MECHANISMS

Mechanisms of harm are not obvious and are not necessarily merely the converse of the intended intervention mechanisms of action. In principle, social interventions might bring about a large range of harms with differing degrees of plausibility. To investigate the most plausible harms and their underlying mechanisms, a priori theorisation is useful. This is essential for quantitative assessment and useful for guiding qualitative

J Epidemiol Community Health: first published as 10.1136/jech-2014-204671 on 17 November 2014. Downloaded from http://jech.bmj.com/ on January 19, 2022 by guest. Protected by copyright.

assessment. Evaluators are increasingly exhorted to develop diagrammatic logic models and descriptive theories of change.[19] These inform the design and conduct of evaluative studies, including the collection of data on the likely causal pathways and the selection of appropriate outcomes.[20] However, the development of logic models usually only focuses on the hypothesised intended beneficial impacts of the intervention. Our suggestion is that complementary models and theories are developed so that they can be used to anticipate the most plausible and most harmful unintended harmful impacts and associated mechanisms. We term these 'dark logic' models. These could enable evaluations to detect both 'harmful externalities' and 'paradoxical effects'. It could also enable evaluations to clarify what mechanisms might underlie detected harms, whether these be paradoxical effects or harmful externalities, in order to produce evidence that might help optimise future interventions and minimise risk of harm.

To develop a dark logic model, evaluators might start by developing their logic model of how the intervention is meant to work. A logic model diagrammatically depicts the inputs that an intervention involves, the processes involved and the mechanisms via which these are intended to realise positive outcomes.[21] In our view, logic models should not simply be linear but should also deal with how intervention mechanisms vary across different contexts.[22] Once the logic for the intervention's intended positive effects has been produced, its assumptions can be scrutinised and the 'dark logic' of potential harms can be constructed. Again, this should address inputs, processes and mechanisms as well as contextual interactions. We recommend using several different approaches to build up a comprehensive dark logic model.

In the first approach, informed by Merton and Giddens,[5 6] the potential mechanisms of intervention harm could be theorised by reflecting on the possible unintended interactions between, on the one hand, the agency (willed actions) of providers, recipients and other stakeholders and, on the other, the social structures that enable and constrain this agency. In the case of social interventions, these structures might relate to the institutions through which the intervention is developed, the manuals guiding how the intervention should be delivered or the resources available for delivery. Structures could also include the wider infrastructure, economic conditions and social norms influencing the broader context in which intervention delivery and receipt will occur. Reflection on how agency and structure might interact in unintended ways might be informed by existing mid-range sociological and psychological theories.

Applying this to the YPDP example, the government specified recruitment and retention targets to manage the programme, but was less strongly focused on targets relating to programme fidelity. Evaluators might have hypothesised a priori that providers would respond creatively to these structural conditions in ways that enabled them to meet their monitoring targets. Evaluators might then have used mid-range sociological theory on the perverse effects of public sector targets[23] to hypothesise that providers' responses to these targets might involve the identification of a captive audience: students who were mandated by their schools to attend the programme on day release instead of their normal schooling (such processes were identified post hoc through qualitative research). The evaluators might then have used mid-range educational theories of labelling[24] to hypothesise that mandating students to miss out on mainstream education to attend programmes might cause them to feel labelled as deviant, producing adverse effects. Thus, in the case of complex interventions such as YPDP, it might be possible that multiple subversions occur acting synergistically to engender harms.

A second approach to theorising harms is to build comparative understanding across similar interventions.[25] This would involve evaluators comparing the logic model of how the intervention being evaluated is meant to work versus the logic models, intervention descriptions and/or process evaluations of similar interventions that have been evaluated previously, practicable only when such an evidence base exists. Ideally, the intervention being evaluated would be compared with some interventions previously evaluated as effective and some previously evaluated as harmful, to illuminate points of corroboration and contradiction. This comparison would be qualitative, drawing in all likelihood on a small number of studies to develop hypotheses, rather than quantitatively testing hypotheses. It is akin to processes of 'reciprocal' and 'refutational' translation which are used to compare and contrast qualitative studies within systematic reviews.[26] Applying this to the YPDP example, evaluators might have compared YPDP with the CAS-Carrera programme, previously reported as effective, which notably differed from YPDP in that it did not target individuals according to whether they were thought to be involved in various risk behaviours.[14] This example of refutational translation might have informed a hypothesis that YPDP targeting would lead to unintended effects via labelling vulnerable students. Evaluators might then have compared YPDP with other interventions that did target young people according to individual risk behaviours and that noted adverse effects arising through processes of labelling and positive deviancy training[2], an example of reciprocal translation.

A third approach to identifying potential harms and underlying mechanisms is to consult with individuals or groups who have particular insights into local contexts and how interventions might operate within these. This is practicable when evaluators have prior access to such stakeholders. Applying this to the YPDP example, evaluators might have conducted early consultations with the managers of agencies charged with delivering YPDP so that their insights could have informed a dark logic model. For example, consultation might have established both that the staff were extremely anxious about achieving recruitment targets and that many were not committed to the YPDP model, seeing it largely as a means of funding existing work. Alongside a consideration of agency/structure interactions and use of mid-range theory, this might have encouraged evaluators to examine how sites varied in their commitment to and faithful delivery of the YPDP model and whether this was associated with potentially harmful mechanisms and effects such as positive deviancy training and increased pregnancies.

IMPLICATIONS OF DARK LOGIC MODELS

We have proposed a method for identifying potential unintended harms. This theorisation of dark logic models is not intended to replace but rather to inform empirical evaluation. Identifying a plausible set of harms and underlying mechanisms would enable evaluators to investigate these using quantitative and qualitative data. Assessing harms empirically may in some cases mean that evaluations require longer periods of follow-up or larger samples depending on the anticipated timescales and prevalence for harms to manifest. Some harms may arise insufficiently frequently to be detected in primary studies but may be assessable via meta-analyses if the harm can be examined consistently across primary studies. Even using our system for pre-hypothesising harms, some unintended harms and the mechanisms that underlie them will remain unanticipated. Thus, our approach does not preclude post hoc investigations about

J Epidemiol Community Health: first published as 10.1136/jech-2014-204671 on 17 November 2014. Downloaded from http://jech.bmj.com/ on January 19, 2022 by guest. Protected by copyright.

how the intervention might have caused harm. We see dark logic models as a means of informing qualitative research but recognise that qualitative research focused on a priori theories should be complemented by grounded qualitative research focused on unanticipated impacts and mechanisms.

Dark logic models might also help programme developers rethink interventions in order that the risk of identified harms might be reduced prior to evaluation. We have given the example of unstructured activities as one aspect of some youth interventions which might be removed or changed to reduce iatrogenic potential. In this way, developing dark logic model is also important for strengthening the interventions we intend to conduct in the first place.

A greater focus on potential harms raises ethical concerns. Evaluators might be expected to inform participants that, though not intended or expected, the possibility of some harms has been anticipated and will be assessed. This should improve the transparency of informed consent. However, it might also cause participants to be more sensitised to potential harms, perhaps correcting previous under-reporting or leading to over-reporting.

CONCLUSION

This paper has suggested that potential iatrogenic effects of public health interventions in the current literature have not been subject to sufficient empirical scrutiny. Social scientists have an ethical obligation to avoid harm beyond merely assessing whether paradoxical effects occurred in the case of intended outcomes. This requires detecting potential harms that might arise from the interventions, but also the mechanisms which could explain these harms, via what we term 'dark logic' models. That is, using prehypothesisation informed by: theorisation of agency-structure interactions; building comparative understanding across similar interventions via reciprocal and refutational translation; and consultation with local actors to identify how mechanisms might be derailed, leading to harmful consequences.

What is already known on this subject

- It is well established that public health programmes involving social or behavioural interventions can sometimes cause serious harms.
- There have been some attempts to categorise the different types of harms of these interventions.
- However, the assessment of adverse effects remains a neglected topic in evaluations of public health interventions.

What this study adds

- A clear argument for the importance of evaluations of public health interventions to detect potential harms and the mechanisms that might underlie these harms so that they might be avoided in future is provided.
- A review of protocols published by the National Institute of Health Research Public Health Research Programme since 2010 suggests that evaluations are inconsistent in their focus on potential harms and very few focus on exploring underlying mechanisms of harms.
- A new systematic process by which evaluators might develop 'dark logic models' to guide the evaluation of potential harms and underlying mechanisms is presented.

Unfortunately, empirical evidence of harms alone is not sufficient to prevent replication. Some interventions consistently shown to be harmful such as 'Scared Straight' programmes continue to be widely delivered.[27] Nonetheless, evidence of harm is a necessary step in preventing the replication of such intervention.

Contributors CB conceived and led the drafting of this paper. FJ, GJM-T and SC contributed to the drafting of this paper.

Competing interests None.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1 Hooker W. *Physician and patient*. New York: Baker and Scribner, 1847.
2 Dishion TJ, McCord J, Poulin F. When interventions harm. *Am Psychol* 1999;54:755–64.
3 Hawe P, Shiell A, Riley T. Theorising interventions as events in systems. *Am J Community Psychol* 2009;43:267–6.
4 Popper K. *The open society and its enemies, Volume 2 Hegel and Marx*. London: Routledge, 1945.
5 Giddens A. *The constitution of society*. Cambridge: Polity Press, 1984.
6 Merton R. The unanticipated consequences of purposive social action. *Am Sociol Rev* 1936;1:894–904.
7 Lorenc T, Oliver K. Adverse effects of public health interventions: a conceptual framework. *J Epidemiol Community Health* 2014;68:288–90.
8 Killeen T, Hien D, Campbell A, *et al*. Adverse events in an integrated trauma-focused intervention for women in community substance abuse treatment. *J Subst Abuse Treat* 2008;35:304–11.
9 Gould MS, Marrocco FA, Kleinman M, *et al*. Evaluating iatrogenic risk of youth suicide screening programs: a randomized controlled trial. *J Am Med Assoc* 2005;293:1635–43.
10 Ogilvie D, Foster CE, Rothnie H, *et al*. Interventions to promote walking: systematic review. *Br Med J* 2007;334:1204.
11 Chou R, Helfand M. Challenges in systematic reviews that assess treatment harms. *Ann Intern Med* 2005;142:1090–9.
12 Lorenc T, Petticrew M, Welch V, *et al*. What types of interventions generate inequalities? Evidence from systematic reviews. *J Epidemiol Community Health* 2013;67:190–3.
13 Smith SW, Hauben M, Aronson JK. Paradoxical and bidirectional drug effects. *Drug Saf* 2012;35:173–89.
14 Philliber S, Kaye JW, Herrling S, *et al*. Preventing pregnancy and improving health care access among teenagers: an evaluation of the Children's Aid Society-Carrera Program. *Perspect Sex Reprod Health* 2002;34:244–51.
15 Wiggins M, Bonell C, Sawtell M, *et al*. Health outcomes of youth development programme in England: prospective matched comparison study. *BMJ* 2009;339:b2534.
16 Imrie J, Stephenson JM, Cowan FM, *et al*. A cognitive behavioural intervention to reduce sexually transmitted infections among gay men: randomised trial. *BMJ* 2001;322:1451–6.
17 Palinkas LA, Atkins CJ, Miller C, *et al*. Social skills training for drug prevention in high-risk female adolescents. *Prev Med* 1996;25:692–701.
18 NIHR. Funded projects. Secondary Funded projects 2014. http://www.nets.nihr.ac.uk/projects/_nocache?collection=netscc&query=&meta_P_sand=project&meta_R_sand=&meta_Q_sand=HTA&status=&start_rank=&sort=&meta_X_prox_or=&meta_S_prox_or=&meta_M_prox_or=&meta_2_sand=&meta_1_sand=&num_ranks=&meta_4_sand=&meta_d=&&programme=PHR&meta_Q_sand=PHR
19 Moore G, Audrey S, Barker M, *et al*. *Process evaluation of complex interventions UK Medical Research Council (MRC) guidance (draft)*. London: Medical Research Council, 2013.
20 Ogilvie D, Cummins S, Petticrew M, *et al*. Assessing the evaluability of complex public health interventions: five questions for researchers, funders, and policymakers. *Milbank Q* 2011;89:206–25.
21 Craig P, Dieppe P, Macintyre S, *et al*. Developing and evaluating complex interventions: the new Medical Research Council guidance. *BMJ* 2008;337:a1655.
22 Pawson R, Tilley N. *Realistic evaluation*. London: Sage, 1997.
23 van Thiel S, Leeuw FL. The performance paradox in the public sector. *Public Perform Manag Rev* 2002;25:267–81.
24 Rist RC. On understanding the processes of schooling: the contributions of labeling theory. In: Karabel J, Halsey AH, eds. *Power and ideology in education*. New York, NY: Oxford University Press, 1977:292–305.
25 Turner S. *Sociological explanation as translation*. New York: Cambridge University Press, 1980.
26 Noblit G, Hare R. *Meta-ethnography: synthesizing qualittive studies*. London: Sage Publications ltd, 1988.
27 Finckenauer JO, Gavin PW. *Scared straight: the panacea phenomenon revisited*. Prospect Heights, Illinois: Waveland Press, 1999.

J Epidemiol Community Health: first published as 10.1136/jech-2014-204671 on 17 November 2014. Downloaded from http://jech.bmj.com/ on January 19, 2022 by guest. Protected by copyright.