

LONDON  
SCHOOL of  
HYGIENE  
& TROPICAL  
MEDICINE



LSHTM Research Online

Petticrew, M; (2003) Why certain systematic reviews reach uncertain conclusions. *BMJ*, 326 (7392). pp. 756-8. ISSN 1468-5833 DOI: <https://doi.org/10.1136/bmj.326.7392.756>

Downloaded from: <http://researchonline.lshtm.ac.uk/8733/>

DOI: <https://doi.org/10.1136/bmj.326.7392.756>

**Usage Guidelines:**

Please refer to usage guidelines at <http://researchonline.lshtm.ac.uk/policies.html> or alternatively contact [researchonline@lshtm.ac.uk](mailto:researchonline@lshtm.ac.uk).

Available under license: Creative Commons Attribution Non-commercial  
<http://creativecommons.org/licenses/by-nc/3.0/>

<https://researchonline.lshtm.ac.uk>

# Education and debate

## Why certain systematic reviews reach uncertain conclusions

Mark Petticrew

The “stainless steel” law of evaluation states that the better designed the outcome evaluation, the less effective the intervention seems. This article explores how this law may be operating in relation to systematic reviews

MRC Social and Public Health Sciences Unit, University of Glasgow, Glasgow G12 8RZ  
Mark Petticrew  
associate director  
mark@msoc.mrc.gla.ac.uk

BMJ 2003;326:756–8

Research syntheses are essential for putting studies in their proper scientific context and are increasingly common in public health, education, crime, and social welfare. A key criticism of systematic reviews, however, is that they are often unable to provide specific guidance on effective (or even ineffective) interventions; instead, they often conclude that little evidence exists to allow the question to be answered. This problem has been recognised in reviews of healthcare interventions,<sup>1</sup> and the electronic journal *Bandolier* recently lamented the absence of systematic reviews containing a solid take home message.<sup>2</sup> However, the problem is even more common in reviews of social and public health interventions, and this paper explains why.

### Sound systematic reviews may not guide practice

In public health there are few trials to review and indeed few other types of outcome assessment.<sup>3</sup> Unsurprisingly, research users often regard reviews of such a limited evidence base as unhelpful and find their conclusions confusing and frustrating.<sup>4</sup> This is ironic, given that systematic reviews are intended (among other things) to reduce uncertainty (box 1). Systematic reviews are certainly capable of doing this, and there are many well known clinical examples.<sup>9</sup> Examples from other fields relevant to public health include two reviews that examined the effectiveness of improved street lighting and

#### Box 1: Systematic reviews and uncertainty

“Systematic reviews aim to reduce uncertainty by strengthening the evidence base”<sup>5</sup>

“Systematic reviews . . . contribute to resolve uncertainty when original research, reviews, and editorials disagree”<sup>6</sup>

“Systematic reviews can be conducted in an effort to resolve conflicting evidence, to answer questions where the answer is uncertain or to explain variations in practice”<sup>7</sup>

“Systematic reviews are needed to inform policy and decision-making about the organisation and delivery of health and social care. They are particularly useful when there is uncertainty regarding the potential benefits or harm of an intervention”<sup>8</sup>

### Summary points

Systematic reviews are often criticised for being unable to provide specific guidance

This is often because the primary studies that they include contain few outcome evaluations

A “stainless steel” law of systematic reviews may also be operating—namely, the more rigorous the review, the less evidence there will be that the intervention is effective

Narrative review methods and narrative and meta-analytic approaches to reviewing observational data need to be improved

Uncertainty will often remain, but systematic reviews help us to acknowledge this and to map the areas of doubt

closed circuit television as deterrents to crime.<sup>10 11</sup> These reviews included a total of 35 studies and found that although closed circuit television reduced crime in car parks, it had little effect in city centres or when used on public transport.<sup>11</sup> Improved street lighting, however, reduced crime by up to a fifth, and savings outweighed the installation costs.<sup>10</sup>

Equally common, however, are reviews that go to extreme lengths to seek out the best evidence, only to conclude that “good evidence is currently lacking.” Although this may be an accurate representation of the state of the evidence, it is not useful for guiding practice or policy, and users and funders will not see value in reviews that consistently and predictably conclude that no good evidence exists. Systematic reviews also risk being perceived, quite wrongly, as simply a means of criticising existing research rather than informing decision making. Worse, their positive messages may be overlooked, and they will be seen as the public health version of Cassandra, the classical bearer of bad news who was doomed never to be believed.

## Too few studies include health outcomes

Sometimes no clear evidence exists simply because the primary studies did not include health outcomes, and in public health in particular the problem often seems to be an absence of evidence rather than evidence of absence of effect. This is partly because in Britain there have been few evaluations of the outcomes of social interventions, including policies, and even fewer have entailed measurement of health outcomes.<sup>3</sup> For example, one of the major uncertainties in public health concerns the health effects of income supplementation (such as changes in taxation or benefits). A recent systematic review found seven trials involving income supplementation, all US based, which examined the impact of a rise of about 14% in people's income; unfortunately, none of the studies had reported reliable data on health outcomes.<sup>12</sup>

## Bricks without straw

Any review starts with defining the question then seeking the appropriate research to answer it. Systematic reviews can be good at answering questions about the effectiveness of specific interventions but often do not yield clear answers to questions about complex interventions that have not themselves been fully evaluated. A review of the evidence can after all only reflect the available primary studies.<sup>13</sup> When outcome evaluations yield little evidence, the range of options for interventions may, however, be informed by expert and other consultations. Qualitative information may give pointers on what is meaningful and acceptable to users; observational evidence (or better, systematic reviews of observational evidence) may show what is potentially effective in the absence of trials; and economic information may show what is affordable. Systematic reviews alone are not a panacea.

## Sifting the evidence

Users of systematic reviews may sometimes suspect that the absence of definite answers is due not to a lack of evidence but to the review process, which typically involves sifting thousands of titles and abstracts for relevance before selecting some—typically less than 20—for in-depth review.<sup>14</sup> Scanning titles and abstracts for relevant studies has some similarities to operating the x ray machines at airports—a life of boredom punctuated by very occasional excitement. The suspicion among non-reviewers may be that among the rejected thousands are many dozens of relevant evaluations that did not meet the review's unreasonably rigorous methodological criteria. Any systematic reviewer will point out, however, that this is not the case. There is generally no hidden pool of relevant studies, qualitative or quantitative, that reviewers are unwilling to include. However large the holes in a reviewer's methodological filter, most research still does not make it through to the other side. Excluded studies are usually rejected on grounds of appropriateness and relevance, rather than on grounds of study design or quality. Quite simply, few relevant outcome evaluations—randomised, controlled, or otherwise—of major UK social programmes have been carried out.



Sifting the evidence for sound studies with a take home message is laborious and the yield disappointing

## The “stainless steel” law of evaluation

It is often said that it is difficult to get answers to “what works” in the case of social interventions because unlike the United States, the United Kingdom has historically had little interest in social experimentation.<sup>15</sup> Not only does Britain lack an experimental culture, it also lacks a strong evaluation culture—at least as far as outcome evaluation of social interventions is concerned. However, even if Britain had an abundance of experimental studies, systematic reviews would still not produce definitive answers. This is because the outcome evaluations on which reviews typically draw are unlikely to identify social “magic bullets.” It has even been suggested that only social programmes that are likely to fail are evaluated; effective programmes are obviously “working” and thus avoid evaluation.<sup>16</sup> Rigorous outcome evaluations of social interventions may therefore be more likely to produce “negative messages,” which may make them unpopular. Oakley has suggested that in the United States, randomised controlled trials of social programmes were funded until they began to show repeatedly negative results, at which point they fell out of favour.<sup>15</sup>

One reason for such negative findings (and by extension for the negative conclusions of many reviews) is that a “stainless steel” law of evaluation may exist. This is one of the “metallic” laws of evaluation drawn up by American sociologist Peter Rossi, derived from a 19th century practice of naming physical laws after substances of varying durability.<sup>16</sup> According to Rossi, the “stainless steel” law states that the better designed the outcome evaluation, the less effective the intervention seems. Rossi also proposed an “iron” law of evaluation, which states that the expected value of any impact assessment of any large scale social programme is zero.<sup>16</sup> The effect is not apparently confined to evaluations of interventions but is also present, for example, in observational epidemiology, where higher risk factor estimates are produced by less rigorous studies.<sup>17</sup> By implication, a stainless steel law of sys-

**Box 2: Definitions**

*Review*—General term for all attempts to synthesise the results and conclusions of two or more publications on a given topic

*Systematic review*—A review that strives to comprehensively identify, track down, and appraise all the literature on a topic (also known as a systematic literature review)

*Meta-analysis*—A review that incorporates a specific statistical strategy for assembling the results of several studies into a single estimate

*Narrative review*—The process of synthesising primary studies and exploring heterogeneity descriptively rather than statistically (that is, by means of a meta-analysis)

tematic reviews also generally applies—that is, the more rigorous the review, the less evidence there will be to suggest that the intervention is effective.

**The low power of the narrative review**

There is another straightforward reason why reviews of social interventions are likely to produce uncertain conclusions. It is because they often use narrative review methods—that is, narratively summarising the results of individual primary studies—and it is rather difficult to detect small intervention effects by this means. The meta-analyst can pool many small studies (all with non-significant results) and by doing so increase the power to detect an effect, thereby reducing the risk of a type II error (false negative result).<sup>18</sup> The narrative reviewer of social interventions often cannot do this, because of substantial heterogeneity in the intervention, outcomes, and context and so is at a greater risk of introducing a type II error (box 2).<sup>14 18</sup>

**Where now?**

Overall, systematic review methods need developing in two main areas. Firstly, the methods of narrative reviews need improving to ensure that reviewers can make effective use of all types of evidence. Secondly, we need to improve the methods of systematic review of observational studies—what Chalmers has referred to as “methodological tiger country.”<sup>18</sup> Advances in both these areas would help to ensure that reviewers can make best use of the available evidence while taking account of heterogeneity in context, study design, and study quality. Uncertainty will always remain, however, particularly when the evidence is unreliable. The singular contribution of systematic reviews in this respect, however, is that they provide reliable maps of these areas of doubt.

**Conclusion**

Systematic reviews do not replace judgment or compassionate reasoning, and absence of clear evidence from systematic reviews does not mean that inertia is the recommended course of action.<sup>19</sup> Lack of clear evidence should not, for example, be a reason for inaction on health inequalities—we should be guided by what we know about the mechanisms by which interventions might plausibly be expected to affect health.<sup>20</sup> After all, at the core of evidence based

decision making is an assumption that decisions may be guided by the best “available” research evidence—and other guidance on action can also be sought.

Black recently stated that the results of single studies are generally not worth disseminating; instead, syntheses of results of studies are the appropriate product of research.<sup>21</sup> Admittedly, such reviews often merely highlight our ignorance, but this in itself is an important contribution. It is, after all, only through mapping what is known and acknowledging uncertainty that scientific knowledge can accumulate. “When you know a thing, to hold that you know it; and when you do not know a thing, to allow that you do not know it—this is knowledge.”<sup>22</sup>

I thank Sally Macintyre, David Ogilvie, and Andy Oxman.

Funding: The author is part of the Economic and Social Research Council's Evidence Network and is funded by the Chief Scientist Office of the Scottish Executive Department of Health.

Competing interests: None declared.

- 1 Alderson P, Roberts I. Should journals publish systematic reviews that find no evidence to guide practice? Examples from injury research. *BMJ* 2000;320:376-7.
- 2 How systematic reviews can disappoint. *Bandolier* 2001 www.jr2.ox.ac.uk/bandolier/ (accessed 10 March 2003).
- 3 Millward L, Kelly M, Nutbeam D. *Public health intervention research: the evidence*. London: Health Development Agency, 2001.
- 4 Nutbeam D. Evidence-based public policy for health: matching research to policy need. *JUHPE Promotion and Education* 2001;2(suppl):15-27.
- 5 Campbell M, Daly C, Wallace S, Cody DJ, Donaldson C, Grant AM, et al. Evidence-based medicine in nephrology: identifying and critically appraising the literature. *Nephrol Dial Transplant* 2000;15:1950-5.
- 6 Mulrow CD, Oxman AD, eds. *Cochrane Collaboration handbook*. In: *Cochrane Library*. Issue 4. Oxford: Update Software, 1997.
- 7 NHS Centre for Reviews and Dissemination. *Undertaking systematic reviews of research on effectiveness. CRD's guidance for those carrying out or commissioning reviews*. York: University of York, 2001. (CRD report No 4, 2nd ed.)
- 8 Egger M, Davey Smith G, O'Rourke K, Rationale, potentials, and promise of systematic reviews. In: Egger M, Davey Smith G, Altman D, eds. *Systematic reviews in health care: meta-analysis in context*. London: BMJ Publishing, 2001.
- 9 Mulrow C. Systematic reviews: rationale for systematic reviews. *BMJ* 1994;309:597-9.
- 10 Farrington D, Welsh B. Improved street lighting and crime prevention. *Justice Quarterly* 2002;19:313-42.
- 11 Welsh B, Farrington D. *Crime prevention effects of closed circuit television: a systematic review*. London: Home Office Research, Development and Statistics Directorate, 2002.
- 12 Connor J, Rogers A, Priest P. Randomised studies of income supplementation: a lost opportunity to assess health outcomes. *J Epidemiol Community Health* 1999;53:725-30.
- 13 Waters E, Doyle J. Evidence-based public health practice: improving the quality and quantity of the evidence. *J Public Health Med* 2002;24:227-9.
- 14 Petticrew M, Song F, Wilson P, Wright K. The DARE database of abstracts of systematic reviews: a summary and analysis. *Int J Technol Assess Health Care* 2000;15:671-8.
- 15 Oakley A. *Experiments in knowing: gender and method in the social sciences*. Cambridge: Polity Press, 2000
- 16 Rossi P. The iron law of evaluation and other metallic rules. *Research in Social Problems and Public Policy* 1987;4:3-20.
- 17 Elvik R. Evaluations of road accident blackspot treatment: a case of the iron law of evaluation studies? *Accid Anal Prev* 1997;29:191-9.
- 18 Egger M, Davey Smith G, Altman D. *Systematic reviews in health care. Meta-analysis in context*. London: BMJ Publishing, 2001.
- 19 Mulrow C, Cook D, eds. *Systematic reviews. Synthesis of best evidence for health care decisions*. Philadelphia: American College of Physicians, 1998.
- 20 Macintyre S. *Evaluating the evidence on measures to reduce inequalities in health*. London: Heath Equity Network, 2002.
- 21 Black N. Evidence based policy: proceed with care. *BMJ* 2001;323:275-9.
- 22 Legge JT. *The analects of Confucius*. Project Gutenberg. Available from <http://gutenberg.net> (Accepted 31 January 2003)

*Endpiece***How many days have you lost?**

A day spent without drawing is a day lost.

Anthony Gormley, British sculptor, born 1950