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.

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, and the electronic journal Bandolier recently lamented the absence of systematic reviews containing a solid take home message.

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. Unsurprisingly, research users often regard reviews of such a limited evidence base as unhelpful and find their conclusions confusing and frustrating. 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.

Examples from other fields relevant to public health include two reviews that examined the effectiveness of improved street lighting and closed circuit television as deterrents to crime. 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. Improved street lighting, however, reduced crime by up to a fifth, and savings outweighed the installation costs.

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. 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.

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. 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. 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 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.

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. 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. 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.

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. 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. 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. By implication, a stainless steel law of sys-
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 stud-
ies are generally not worth disseminating; instead, syn-
theses of results of studies are the appropriate product
of research.11 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 uncer-
tainty 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.”12
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.

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).13 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).14 15

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.”16 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.17 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.18 After all, at the core of evidence based

Box 2: Definitions

\[\text{Review—General term for all attempts to synthesize the results and conclusions of two or more publications on a given topic.}\]

\[\text{Systematic review—A review that strives to comprehensively identify, track down, and appraise all the literature on a topic (also known as a systematic literature review).}\]

\[\text{Meta-analysis—A review that incorporates a specific statistical strategy for assembling the results of several studies into a single estimate.}\]

\[\text{Narrative review—The process of synthesizing primary studies and exploring heterogeneity descriptively rather than statistically (that is, by means of a meta-analysis).}\]

3 Millboard L, Kelly M, Nutbeam D. Public health intervention research: the evi-
4 Nutbeam D. Evidence-based public policy for health: matching research to policy need. JUHPE; Promotion and Education 2001;2(suppl):15-27.
12 Cooper J, Rogers A, Priest P. Randomised studies of income supplemen-

Endpiece

How many days have you lost?
A day spent without drawing is a day lost.
Anthony Gormley, British sculptor, born 1950