Systematic reviews of health effects of social interventions: 2. Best available evidence: how low should you go?

David Ogilvie, Matt Egan, Val Hamilton, Mark Petticrew

Study objective: There is little guidance on how to select the best available evidence of health effects of social interventions. The aim of this paper was to assess the implications of setting particular inclusion criteria for evidence synthesis.

Design: Analysis of all relevant studies for one systematic review, followed by sensitivity analysis of the effects of selecting studies based on a two dimensional hierarchy of study design and study population.

Setting: Case study of a systematic review of the effectiveness of interventions in promoting a population shift from using cars towards walking and cycling.

Main results: The distribution of available evidence was skewed. Population level interventions were less likely than individual level interventions to have been studied using the most rigorous study designs; nearly all of the population level evidence would have been missed if only randomised controlled trials had been included. Examining the studies that were excluded did not change the overall conclusions about effectiveness, but did identify additional categories of intervention such as health walks and parking charges that merit further research, and provided evidence to challenge assumptions about the actual effects of progressive urban transport policies.

Conclusions: Unthinking adherence to a hierarchy of study design as a means of selecting studies may reduce the value of evidence synthesis and reinforce an "inverse evidence law" whereby the least is known about the effects of interventions most likely to influence whole populations. Producing generalisable estimates of effect sizes is only one possible objective of evidence synthesis. Mapping the available evidence and uncertainty about effects may also be important.

Despite increasing calls for systematic reviews of health effects of social interventions, there is little methodological research or even guidance on how such reviews should be done. We have lifted the lid on the "private life" of the input side of one such systematic review to expose some of our methodological processes and decisions to critical analysis. In a companion paper, we set the scene and examined one phase of the review, the search for evidence. In this paper, we examine another phase of the review: the selection of evidence for inclusion. We investigate the effect of varying our inclusion criteria on the findings and overall value of the review.

SELECTING EVIDENCE FOR INCLUSION

Researchers designing systematic reviews of intervention studies are advised to specify their research questions in terms of four facets: the intervention, the population receiving the intervention, the outcome of interest, and the study designs deemed worthy of inclusion. This approach is undoubtedly helpful for structuring research questions and protocols, but we aimed to synthesise population level evidence in a cross disciplinary field where comparatively little empirical intervention research has been done. A broad understanding of population health and its wider determinants implied a need to frame our primary research question rather differently. We were not asking, for example, "What is the evidence that traffic calming leads to a change in travel behaviour?", but rather "What interventions, of any kind, lead to such a change?". In other words, we focused on the outcome of interest and were open to the possibility that any kind of intervention might contribute towards achieving it. This is an example of addressing a "broad" review question—acknowledged as a valid, but often difficult, type of review to carry out. Broad questions are also often appropriate in other types of evidence synthesis, such as that used in health impact assessment.

Many published systematic reviews have only considered evidence from randomised controlled trials (RCTs). The motivation is to minimise bias, but the Cochrane handbook recognises that this can compromise the relevance of a review, and asks (but does not answer) the question "How far is it possible to achieve a higher level of relevance by including evidence other than that derived from RCTs without violating the central principle: minimising bias?" There are already precedents for varying the inclusion criteria for study design according to the nature of the available evidence. For example, although some reviews published by the Cochrane Tobacco Addiction Group are restricted to RCTs, those on community or population level interventions include other study designs including, in some cases, uncontrolled before and after studies.

It is increasingly recognised that the usual approach to selecting studies based on a "hierarchy of evidence" may rely too heavily on study design as a marker of validity or utility. This may favour interventions most amenable to certain types of study design, particularly those with a medical rather than a social focus and those that target individual people rather than populations. This type of bias has been described as "methodological imperialism" that could distort, rather than strengthen, the evidence base.

The relative lack of methodological research on how to deal with evidence from studies other than RCTs may make researchers feel vulnerable at key decision points in the process of synthesising evidence. In this paper we describe how we selected studies for inclusion. We then analyse the utility of the different types of studies identified, report a
We have reported details of our methods previously. Briefly, can be expected to contribute to the evidence base.

**METHODS**

We have reported details of our methods previously. Briefly, we designed a wide search strategy defined entirely in terms of the outcome of interest. We screened the titles and abstracts, examined the full text of any documents that appeared relevant, and finally identified 69 relevant studies that met our preliminary criteria (see box).

We carried out full data extraction and critical appraisal on all of these relevant studies, and therefore formed an overview of the full range of study populations, interventions, and study designs available in the field as well as the range of outcome metrics used and effect sizes identified.

It became clear that both the types of study design and the nature of the study populations varied widely. Some studies had used comparatively robust methods to measure, for example, changes in vehicle flows along certain roads, but these studies could tell us nothing about the people using those vehicles or about their non-vehicular (walking) trips. Similarly, we found studies showing how the distribution of transport mode choice had changed among weekend shoppers interviewed in a city centre street, but these studies could tell us nothing about where the shoppers had come from or whether their overall travel behaviour had changed.

We also found particular difficulty in deciding what to do with articles—typically book chapters—about “successful” towns or cities in which trends in travel patterns were linked post hoc to a variety of interventions, often part of a complex integrated urban policy that included land use planning, public transport improvements, widespread traffic restraint, cycle routes, pedestrianisation, and related measures. These articles did not seem to report the results of specific studies of specific interventions as such, so we characterised them as “case studies” in which authors had reported trends of interest to us, but had not presented data in a way that enabled us to assess the strength of the causal assertions being made.

These findings led us to devise a simple matrix, or two dimensional hierarchy, of study utility (table 1). We categorised studies not only on the study design (a marker of internal validity) but also on the study population, which we took as our primary marker of external validity—in other words, a marker of how useful the study would be for answering our question about changes in population health and health determinants. We plotted the distribution of all relevant studies in this matrix and used it to specify our final inclusion criteria. We further assessed and summarised the internal validity of included studies using 10 methodological criteria.

When our review was complete, we also conducted a sensitivity analysis to examine what the content and findings of the review would have been if we had taken one of two extreme approaches to inclusion—either (a) by restricting the review to randomised controlled trials, or (b) by including all relevant studies. This sensitivity analysis was intended to answer two questions: were the conclusions of our review sensitive to the inclusion criteria, and could we have reached our conclusions more efficiently?

**RESULTS**

**Two dimensional hierarchy**

We examined the distribution of studies in the matrix (fig 1) and chose final thresholds for inclusion. These were, of course, still somewhat arbitrary but were based on having reviewed all available relevant studies in detail.

We first excluded studies whose design was neither prospective nor controlled (n = 29). We then excluded studies whose populations did not represent a local population or subset thereof (n = 9). This left 31 studies (represented by the dark columns in the figure). Nine of these were subsequently excluded on the grounds that they contained inadequate information about methods, results or both, leaving 22 studies finally included in the review.

**Sensitivity analysis**

**Effect of including only RCTs**

We found only three RCTs. If we had included only these studies, we would have benefited from reviewing a small set of studies that were well written and comparatively easy to appraise. These were also the only studies that contained robust data on direct health outcomes. However, we would only have been able to include evidence about two small categories of intervention: targeted behaviour change programmes for commuters, and school travel coordinators. We would not have identified any evidence about, or perhaps even the existence of, any population wide health promotion activities, “environmental” engineering or transport service developments, or financial incentives, and we would not have

| Table 1 Two dimensional hierarchy of study utility |
|---------------------------------|---------------------------------|
| Study design | Study population |
| Randomised controlled trial | Households or local residents |
| Controlled panel study (repeated measures on the | Subset of local population (drivers, commuters, or school |
| same participants) | pupils) |
| Controlled repeated cross sectional study | Participants in a targeted intervention already selected |
| Controlled retrospective study | from one of the groups above |
| Uncontrolled panel study (repeated measures on the | Passers by at, or visitors to, a study location |
| same participants) | Patients* |
| Uncontrolled repeated cross sectional study | Vehicles |
| Uncontrolled retrospective study | Population not clear |
| Case study of trends in mode share | |
| Design not clear | |

The effectiveness of interventions given in a clinical setting was outside the scope of the review.
identified any of the studies that indicated possible unex-
pected or inequitable effects of interventions.\textsuperscript{13}

Evidence provided by excluded studies
We identified several types of evidence provided by studies
we did exclude, which are summarised in table 2, grouped by
type of intervention.

A larger taxonomy of interventions of interest
Some specific types of intervention were only represented in
excluded studies: health walks, parking charges, and fuel
rationing. Most of these studies indicated potential for a
positive effect, albeit based on designs with important
methodological weaknesses with respect to our review
question. These types of intervention therefore merit more
detailed consideration by researchers and policymakers.

Evidence about some interventions consistent with the
stronger evidence already included in the review
We had found the strongest evidence of positive effects in the
area of targeted behaviour change programmes (based on six
studies of four interventions).\textsuperscript{13} Two excluded studies of
targeted programmes also identified potential for positive
effects, as did two excluded studies of workplace schemes
involving free bikes. We also found a large number of
excluded studies of engineering measures whose findings
were broadly consistent with our primary finding of little
evidence of positive effects, and single excluded studies of
road user charging and alternative transport services that did
not contradict our primary findings.

Evidence about one category of intervention that could
contradict our primary findings
We excluded two studies of publicity campaigns for sustain-
able transport that both claimed a substantial positive effect.
Neither study was reported in sufficient detail for our
purposes (for example, there were no details of sampling
method, response rate, survey instrument, and so on), we
could not find any more detailed reports, and authors did not
reply to a request for more information. It is therefore
possible that evidence exists to contradict our primary
finding of little evidence of effectiveness for publicity
campaigns, although it seems unlikely that such evidence
would be strong.

Evidence to challenge assumptions about “successful”
cities
Even if it were possible to attribute the observed trends in
travel patterns in “case study” cities to part or all of their
multifaceted urban transport policies, a positive change (in
our terms) was only actually reported in three of the 13 cities,
and in two of these that positive change was only seen for
trips into the city centre and not for residents’ trips overall.
Where modal shifts were reported, these were more likely to
be, for example, an increase in public transport at the
expense of all other modes including walking and cycling.

DISCUSSION
Hierarchies of evidence for public health
We reported previously that the most robust evidence of
effectiveness was concentrated around interventions targeted
on motivated groups of volunteers.\textsuperscript{13} Our subsequent analysis
shows that this “evidence bias” may reflect, at least partly, an
“evaluative bias”: other types of intervention (especially
those applied to whole populations or areas) have tended to
be evaluated using less rigorous methods. For those inter-
ested in improving population health, the most useful
evidence is likely to come from population level studies with
designs of high internal validity—those located in the far

What is already known on this subject?

- We need better syntheses of evidence about the effects
  of interventions to influence the wider determinants of
  health
- Some have questioned whether selecting evidence
  according to a rigid, unidimensional hierarchy based
  on study design—for example, only including ran-
  domised controlled trials—is appropriate in this field
- We lack an accepted, evidence based methodology for
  selecting useful evidence for inclusion in evidence
  synthesis.
Table 2  Evidence from excluded studies

<table>
<thead>
<tr>
<th>Category of intervention</th>
<th>Included studies</th>
<th>Summary of findings from included studies*</th>
<th>Excluded studies</th>
<th>Primary reasons for exclusion</th>
<th>Summary of findings from excluded studies*</th>
<th>Comparison with included studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Targeted behaviour change programmes</td>
<td>One randomised trial, three controlled prospective studies of the same intervention in different settings, and two uncontrolled prospective studies</td>
<td>Modal shift in motivated subgroups (five out of six studies)</td>
<td>One uncontrolled prospective study of advising patients to take more exercise, one pilot study of targeted information for households and commuters</td>
<td>Population</td>
<td>Both identified potential for modal shift</td>
<td>Consistent</td>
</tr>
<tr>
<td>Agents of change and publicity campaigns</td>
<td>One randomised controlled trial of school travel coordinators, one controlled prospective and two uncontrolled prospective studies of publicity campaigns</td>
<td>Small modal shift in only one relatively weak study out of four</td>
<td>One case study of a travel management association, two studies of sustainable transport campaigns, one uncontrolled prospective and one controlled retrospective</td>
<td>Design</td>
<td>“Inability to achieve any significant shift in travel behaviour”</td>
<td>Consistent</td>
</tr>
<tr>
<td>Health walks</td>
<td>No included studies</td>
<td>—</td>
<td>Two uncontrolled retrospective studies of participants</td>
<td>Design</td>
<td>About a quarter of participants claimed to have changed their travel behaviour</td>
<td>Additional category of intervention with potential for positive effect—requires further research</td>
</tr>
<tr>
<td>Cycling promotion</td>
<td>One uncontrolled prospective study (also included in the group of targeted behaviour change programmes above)</td>
<td>Intensive targeted programme including free bike produced a modal shift</td>
<td>Two uncontrolled retrospective studies of free workplace bikes plus local infrastructure improvements</td>
<td>Design</td>
<td>Both identified the potential for modal shift</td>
<td>Consistent</td>
</tr>
<tr>
<td>Networks of routes for cyclists and pedestrians</td>
<td>One controlled prospective study and two uncontrolled prospective studies</td>
<td>Increase in cycling mode share (only) in only one of three studies</td>
<td>Five studies, all either case studies reported with scant detail or based on vehicle counts</td>
<td>Design (4)</td>
<td>One study based on vehicle counts reported increases in cycling, but no data on walking; the others showed no evidence of modal shift</td>
<td>Additional category of intervention with no evidence of positive effect</td>
</tr>
<tr>
<td>Traffic restraint</td>
<td>Three uncontrolled prospective studies: one of city centre traffic restraint, one of bypasses and one of 20 mph (30 km/h) zones</td>
<td>Small modal shift in only one of three studies</td>
<td>One uncontrolled retrospective study of a variety of urban traffic calming schemes, either using unclear or case study designs or based on vehicle counts or shoppers, seven studies of a variety of urban traffic restraint schemes, either using unclear or case study designs or based on vehicle counts or shoppers</td>
<td>Design</td>
<td>Small proportions of residents claimed to have changed their travel behaviour</td>
<td>Consistent overall</td>
</tr>
<tr>
<td>Road user charging</td>
<td>One uncontrolled prospective study</td>
<td>No evidence of modal shift</td>
<td>One pilot study of the effect of an in-car charging unit on commuting journeys</td>
<td>Information</td>
<td>“Another peak-period alternative for two of the sample was cycling”</td>
<td>Consistent—no clear quantification of modal shift</td>
</tr>
<tr>
<td>Parking charges</td>
<td>No included studies</td>
<td>—</td>
<td>One uncontrolled prospective study of town centre parking charges</td>
<td>Information</td>
<td>Identified potential for modal shift among commuters, but not among residents</td>
<td>Additional category of intervention with potential for positive effect—requires further research</td>
</tr>
</tbody>
</table>
### What does this study add?

- Relying on randomised controlled trials would have seriously compromised the scope and value of our evidence synthesis.
- Relevant, population level evidence is dispersed across a wide range of types of study; mapping all of this evidence is a useful exercise in its own right and may be an important part of the process of selecting the most useful evidence for final inclusion.
- Filtering out studies for exclusion without examining them in detail may deprive both reviewers and users of important evidence and insights.

## Table 2

<table>
<thead>
<tr>
<th>Category of intervention</th>
<th>Included studies</th>
<th>Excluded studies</th>
<th>Relevant, population level evidence is dispersed across a wide range of types of study; mapping all of this evidence is a useful exercise in its own right and may be an important part of the process of selecting the most useful evidence for final inclusion.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fuel rationing</td>
<td>One controlled prospective study of a new railway station</td>
<td>Not designed to collect relevant outcome data</td>
<td>Additional category of intervention with potential for positive effect — requires quantification of modal shift.</td>
</tr>
<tr>
<td>Providing alternative services</td>
<td>One controlled prospective study of a new railway station</td>
<td>Not designed to collect relevant outcome data</td>
<td>Additional category of intervention with potential for positive effect — requires quantification of modal shift.</td>
</tr>
<tr>
<td>Multimodal urban transport policies</td>
<td>One controlled prospective study of a new railway station</td>
<td>Not designed to collect relevant outcome data</td>
<td>Additional category of intervention with potential for positive effect — requires quantification of modal shift.</td>
</tr>
</tbody>
</table>

---

Our findings therefore support concerns raised elsewhere that rigid or simplistic adherence to a hierarchy of study design as the primary marker of study utility may be unhelpful, particularly in the fields of health promotion and public health.\(^{4,5}\) For example, the interventions studied in RCTs represent only a small subset of all those that could be or have been adopted. We support the use of RCTs where possible, but many interventions of interest in public health cannot be studied in this way for scientific, political, or practical reasons.\(^{4,6}\) Extending the inclusion criteria as far as we did enabled us to review evidence about a much larger range of interventions and identify some pointers towards potential unexpected effects.\(^{4,7}\) Having re-examined the evidence contained in the studies we did exclude, we do not think that we unwittingly censored any convincing evidence of effectiveness. However, we did identify some interventions that could have positive effects and should be the subject of further research. We also identified other studies, notably the case studies of cities frequently cited as examples of good practice in transport policy, in which there was no actual evidence of success in promoting walking and cycling as an alternative to using cars.

### Best available evidence

The Cochrane handbook acknowledges a place for systematic reviews that address broad questions, but warns of potential difficulties with synthesising and interpreting data from a large set of heterogeneous studies. Identifying all relevant studies is part of what distinguishes a systematic review from a traditional narrative review.\(^{4,8}\) We developed our inclusion criteria iteratively by searching widely, fully appraising all relevant studies, and thereby forming an overview of all available evidence before deciding what should be included.\(^{9}\) Others have also acknowledged that it may not always be possible to specify inclusion criteria in advance\(^{10}\) and that the definition of “relevant” studies may emerge through an extended process of searching, scanning, production of criteria, and further searching.\(^{11}\)

Our approach reflects the principle described by Slavin as “best evidence synthesis”, in other words, not allowing a desire for the “best” evidence to stand in the way of using the...
best available evidence. In a review of the effectiveness of strategies for transferring patient information, Badger et al. framed the reviewer’s task as to review and evaluate “such research as is available.” This did not mean they abandoned the need for critical appraisal; rather, they made informed judgments about the utility of different studies in the light of the whole range of studies available.

What is evidence synthesis for?

The answer to the question “How low should you go?” depends on what researchers think evidence synthesis is for and what evidence is available in a given topic area. Evidence synthesis is often undertaken with the objective of pooling results to produce generalisable estimates of effect size, preferably (in some circles) using the formal technique of meta-analysis. We found that the “best available evidence” in our topic area did not permit us to do this. Is such an objective necessary for a systematic review of intervention studies? A recent editorial highlighted disagreement between authors and peer reviewers over whether the topic of a systematic review of community based interventions was sufficiently coherent or precise to permit generalisation, and argued that learning in public health is best promoted by the critical sharing of evidence, not by censoring suboptimal evidence. Systematic reviews may contribute to public health decision making in various ways. Hammersley has argued that “synthesis” may mean different things to different people, identifying one particular use of the word common among qualitative investigators but not systematic reviewers: producing a mosaic or map in which the distinctive, complementary contributions from different studies are combined to produce a “bigger picture.” This meaning, which is in sharp contrast with the pooling of data from homogeneous studies in a meta-analysis, perhaps reflects more closely what our review achieved. One aspect of this “bigger picture” is the articulation of uncertainty—about the effectiveness of interventions, about the research undertaken on them, and about their potential for unexpected or inequitable effects. Alderson has argued that we should not be embarrassed to admit uncertainty, but should admit it so that the evidence base can then be strengthened. We do not, of course, suggest that reviewers should incorporate the results of less robust studies but not systematic reviewers: producing a mosaic or map in which the distinctive, complementary contributions from different studies are combined to produce a “bigger picture.”

Is the systematic review a fraud?

Handbooks and protocols for systematic reviews, and the reports of their findings, can often given the impression of a linear, rational research process driven by a set of decisions made a priori. But the further a review strays from the world of the placebo controlled drug trial, the less tenable this idea becomes. In this respect, a report of a systematic review is no different from any other scientific publication: it can give a misleading narrative of the research process. The evidence never speaks for itself, but is always open to interpretation, and there are elements of the review process that entail judgment and cannot be made entirely transparent or replicable. Designing and conducting systematic reviews of the health effects of interventions to influence the wider determinants of health is a difficult task for which a standard methodology—whether for searching, study selection, or any other part of the process—has not yet emerged. The methods we have adopted, and our decision to scrutinise them, are open to challenge. None the less, we suggest that it is preferable to reach conclusions, however tentative, that are based on the best available evidence rather than simply stating that no evidence is available.

Authors’ affiliations

D Ogilvie, M Egan, M Petticrew, MRC Social and Public Health Sciences Unit, University of Glasgow, UK

V Hamilton, Development and Alumni Office, University of Glasgow

Funding: the review was funded by the Chief Scientist Office of the Scottish Executive Health Department and by the ESRC Evidence Network. DO is now funded by a Medical Research Council fellowship. The funding sources played no part in the design, analysis, interpretation, or writing up of the study or in the decision to publish.

Competing interests: none known.

Ethical approval: not required.

A list of references to the studies excluded from the systematic review is available on request from the first author.

REFERENCES


Clinical Evidence—Call for contributors

Clinical Evidence is a regularly updated evidence-based journal available worldwide both as a paper version and on the internet. Clinical Evidence needs to recruit a number of new contributors. Contributors are healthcare professionals or epidemiologists with experience in evidence-based medicine and the ability to write in a concise and structured way.

Areas for which we are currently seeking authors:
- Child health: nocturnal enuresis
- Eye disorders: bacterial conjunctivitis
- Male health: prostate cancer (metastatic)
- Women’s health: pre-menstrual syndrome; pyelonephritis in non-pregnant women

However, we are always looking for others, so do not let this list discourage you.

Being a contributor involves:
- Selecting from a validated, screened search (performed by in-house Information Specialists) epidemiologically sound studies for inclusion.
- Documenting your decisions about which studies to include on an inclusion and exclusion form, which we keep on file.
- Writing the text to a highly structured template (about 1500–3000 words), using evidence from the final studies chosen, within 8–10 weeks of receiving the literature search.
- Working with Clinical Evidence editors to ensure that the final text meets epidemiological and style standards.
- Updating the text every six months using any new, sound evidence that becomes available.
- To expand the topic to include a new question about once every 12–18 months.

If you would like to become a contributor for Clinical Evidence or require more information about what this involves please send your contact details and a copy of your CV, clearly stating the clinical area you are interested in, to Klara Brunnhuber (kbrunnhuber@bmjgroup.com).

Call for peer reviewers

Clinical Evidence also needs to recruit a number of new peer reviewers specifically with an interest in the clinical areas stated above, and also others related to general practice. Peer reviewers are healthcare professionals or epidemiologists with experience in evidence-based medicine. As a peer reviewer you would be asked for your views on the clinical relevance, validity, and accessibility of specific topics within the journal, and their usefulness to the intended audience (international generalists and healthcare professionals, possibly with limited statistical knowledge). Topics are usually 1500–3000 words in length and we would ask you to review between 2–5 topics per year. The peer review process takes place throughout the year, and our turnaround time for each review is ideally 10–14 days.

If you are interested in becoming a peer reviewer for Clinical Evidence, please complete the peer review questionnaire at www.clinicalevidence.com or contact Klara Brunnhuber (kbrunnhuber@bmjgroup.com).