Operational criteria of causality for observational road safety evaluation studies

Rune Elvik

Institute of Transport Economics, Gaustadalléen 21, NO-0349 Oslo, Norway
(E-mail: re@toi.no)

ABSTRACT

This paper discusses nine criteria that can be used to assess whether the effects attributed to road safety treatments in studies evaluating the effects of such treatments are causal or merely reflect randomness, confounding or non-causal statistical association. The criteria are to a major extent based on similar criteria developed in epidemiology to assess the causality of factors associated with diseases. The criteria are related to:

1. The presence of a statistical association between a treatment and (changes in) road safety (effect).
2. The strength of the statistical association between treatment and effect.
3. The internal consistency of the statistical association between treatment and effect.
4. The clarity of the causal direction between treatment and effect.
5. The invariance of the statistical association between treatment and effect with respect to confounding factors.
6. The identification and measurement of causal mechanisms generating the association between treatment and effect.
7. The possibility of giving a theoretical explanation of the association between treatment and effect.
8. The presence of a dose-response pattern in the association between treatment and effect.
9. The specificity of the association between treatment and effect to the intended target group of the treatment.

These criteria can be used to assess the quality of road safety evaluation studies. The more criteria a study satisfies, the better is its quality. Use of the criteria is illustrated by means of examples.

Key words: Causality, evaluation study, road safety, study quality

INTRODUCTION

A successful randomised controlled trial is generally regarded as the only study design that permits causal inferences, i.e. claims that the effects found were caused by the treatment applied, and not something else. In the field of road safety, randomised controlled trials are rare. It is therefore relevant to ask how causality can be assessed in observational road safety evaluation studies. Is it at all possible? Epidemiologists have faced the same question in trying to determine the causes of disease. Criteria of causality have been developed to help answer questions such as: does smoking cause lung cancer? Based on the criteria of causality developed in epidemiology (1-5), this paper will discuss the following research problems:
1. Can criteria of causality for road safety evaluation studies be developed, taking the criteria developed in epidemiology as a model?

2. Provided criteria of causality can be developed, how conclusive are these criteria in supporting a causal interpretation of road safety evaluation studies?

The main objective of the paper is to put forward a list of criteria for discussion and further development. Use of the criteria will be illustrated by means of two case studies that may hopefully serve as a basis for replication to provide a broader basis for judging the applicability of the criteria.

CRITERIA OF CAUSALITY

Before listing the criteria of causality, it is perhaps useful to define causality formally: a road safety treatment causes changes in road safety if it produces changes in road safety that would otherwise not have occurred. To produce a change is to act as a necessary condition for the change. Without the safety treatment, the change would not have occurred. Nine criteria of causality for observational road safety evaluation studies (i.e. studies designed to evaluate the effect of safety treatments) are proposed. These criteria are intended to be applicable both to a single study and to a set of studies.

1. There should be a statistical association between treatment and variables measuring its effects (e.g. number of accidents, accident rate, number of injured road users).

2. A strong statistical association between treatment and effect is more likely to be causal than a weak statistical association.

3. The statistical association between treatment and effect should be internally consistent, i.e. identical within the bounds of randomness in all subsets of data.

4. The direction of causality between treatment and effect should be clear, i.e. it should be clear that the treatment is (one of) the cause(s) of the effect and not the other way around.

5. The association between treatment and effect should not vanish when potentially confounding factors are controlled for.

6. A causal mechanism generating the association between treatment and effect should be identified and described statistically.

7. It should be possible to account for the association between treatment and effect in theoretical terms.

8. There should be a dose-response pattern in the relationship between treatment and effect (provided the treatment comes in different doses).

9. The effect should only be found within the target group of the treatment.

The first criterion basically means that there should be an effect, i.e. changes in variables measuring road safety associated with the introduction of the safety treatment should be observed. If there is no change in the expected number of accidents (or other dependent variables), it is in principle still possible that the treatment has had an effect. In such a case, however, the effect of the safety treatment has been offset by the effects of other factors. This is one of the reasons why the strength of the statistical association between treatment and effect is often proposed as a criterion of causality. If the effects of all other factors are stronger than the effects of the treatment, it may be unable to produce the changes that are the hallmark of causality.

Consistency has for a long time been regarded as one of the distinguishing characteristics of a causal relationship (6). This has probably to do with the fact that criteria of causality were first developed for the natural sciences, in which law-like relationships are observed more
often than in social science. Every time you heat an iron rod, it will expand – and by nearly the same amount for the same temperature. It does not contract 30% of the time and expand the other 70%. According to theories of scientific explanation in the natural sciences (7), the so called “covering law” model of explanation is common in these sciences. This means that statements describing causal relationships should, ideally speaking, be statements of scientific laws.

The social sciences are, of course, entirely different. In fact, they are so different that in some branches of social science – history is a case in point – the existence of law-like relationships is denied. Historians have tried to identify the causes of revolutions or civil war. They are unable to come up with any general causes. All that can be said is that when conditions A, B, C ... etc are found, then sometimes there will be civil war and sometimes there will not.

Hauer (8), in discussing cause and effect in observational cross-section studies on road safety, also emphasises consistency in the findings of a single study or several studies as an important criterion of causality. However, this does not mean that variation in estimates of effect as such invalidates causality. The point is that the variation should be explicable in terms of either: (a) a dose-response pattern, (b) a specificity of an effect to a target group, or (c) a known causal mechanism – or combinations of these (see below). If we observe inexplicable variation in effects, this casts doubt on causality, since a given cause ought to produce a given effect, at least within the bounds of random variation.

Causal relationships can go in both directions, in the sense that A may be a cause of B and B in turn a cause of A. Homeostatic processes, like the regulation of body temperature can be thought of as bi-directional causation. As far as road safety is concerned, however, we are searching for causal relationships that go in one direction only: from treatment to safety. The cause normally comes before the effect in time. To determine causal direction means to ascertain that the treatment was the cause of changes in safety, and not the other way around. A well-known case of changes in safety causing a treatment is when an abnormally high recorded number of accidents leads to the introduction of a treatment. When the number of accidents subsequently goes down, this may be the result of regression-to-the-mean only.

This case can also be treated as a case of incomplete control for confounding. In general, confounding refers to any alternative explanation for the observed changes in safety. Thus, to rule out confounding, it is necessary to remove (control for) the effects of any potentially confounding factor. Only if the effects attributed to the safety treatment still persist can they be regarded as causal. Obtaining adequate control for potentially confounding factors is perhaps the most difficult criterion of causality to satisfy in observational road safety evaluation studies.

Few, if any, road safety treatments affect safety directly. Such treatments influence safety by modifying one or more risk factors that are associated with accidents. Thus, the causal mechanism through which safety treatments generate their effects is the modification of risk factors that affect accident rate or injury severity. It is important to note, however, that many safety treatments influence not just the risk factors they are intended to influence, but also other risk factors (9). This influence generally takes the form of behavioural adaptation on the part of road users. A complete description of the causal mechanism of a road safety treatment should therefore identify and measure changes in both the target risk factors and those risk factors that represent any behavioural adaptation to the treatment.

As already mentioned, scientific explanations in the natural sciences are characterised by reference to scientific laws or general theories. One of the problems of road safety evaluation studies is that few findings can be ruled out on theoretical grounds. Yet, one may at least treat some findings as less plausible than others by reference to theory, in particular laws of physics and perception. One would normally expect lower speed to be associated
with fewer and less serious accidents (10). While contrary findings cannot be ruled out, it would be more difficult to invoke laws of physics to explain why less kinetic energy should lead to more accidents and more serious injuries than the opposite.

A dose-response pattern between cause and effect is relevant for road safety evaluation studies in two cases. One case is when the safety treatment itself comes in different doses, such as different amounts of training, different levels and intensities of police enforcement, different frequencies of technical inspections, etc. The other case is when the size of the effect of a safety treatment on its target risk factor (or factors) varies within the sample studied. One would then, all else equal, expect the largest effects on accidents or injuries to be associated with the largest changes in the target risk factors.

A related source of variation in the size of the effects of a safety treatment is moderator variables. A moderator variable can perhaps be treated as a target risk factor for a safety treatment, but in some cases moderator variables are not directly influenced by the safety treatment. A case in point would be the length and complexity of the main road through a town when a bypass is built around it. This road may remain unchanged when the bypass is built; yet one would expect a bypass road to have greater effect when built to bypass a long and complex urban road than when built to bypass a short and less complex urban road.

Some safety treatments have clearly defined target groups, for example young motorcyclists riding motorcycles with a large engine. When young motorcyclists are restricted to smaller machines, one would expect the number of accidents to go down for young motorcyclists riding large machines, but not in other groups of road users. For a case in point, see Broughton (11).

It should be noted that criteria 8 and 9 – dose-response relationship and specificity of an effect to a clearly defined target group – are optional. These criteria can only be applied if a meaningful dose-response pattern can be identified or if there is a clearly designated target group for a treatment.

**OPERATIONAL DEFINITIONS OF THE CRITERIA**

Operational definitions of the criteria are provided in Table 1 and are briefly reviewed here.

There are multiple indicators of statistical association between a road safety treatment and its effects. Some studies use simple odds or odds ratios to measure effects, other studies use regression coefficients or correlation coefficients. Whatever the indicator of effect used, it should first of all be statistically significant, i.e. indicating a change in safety greater than random variation. Thus, criterion 1 is operationally defined in terms of the existence of a statistically significant change in safety associated with the safety treatment whose effects a study is designed to evaluate. Secondly, it is proposed to measure the strength of an association (criterion 2) by means of the effect size. Effect sizes are commonly used in meta-analysis (12) to make effects stated in different metrics comparable. An effect size is the conversion of an estimate of effect to a Z-score, i.e., stating an effect in terms of the number of standard deviations it differs from zero:

\[
\text{Effect size} = \frac{\text{Best estimate}}{\text{Standard error of estimate}}
\]

Thirdly, the consistency of several estimates of effect (criterion 3) can be assessed in terms of the overlap of their confidence intervals, as suggested by Elvik (13). Thus, consider the following three 95% confidence intervals for different estimates of effect:

Interval 1: (-19%, +7%); spanning 26 percentage points
Interval 2: (-11%, -3%); spanning 8 percentage points
Interval 3: (-22%, -7%); spanning 15 percentage points

Three pair-wise comparisons (1 vs 2, 1 vs 3, 2 vs 3) can be made on the basis of these confidence intervals. Comparing confidence intervals 1 and 2, it can be seen that they are consistent with respect to direction and magnitude of impact. The smaller confidence interval (2) is entirely contained within the larger (1). Confidence intervals 1 and 3 are also consistent with respect to direction and magnitude of impact, but do not fully overlap. A consistency score can be assigned as a function of the degree of overlap between two confidence intervals. Interval 1 is the larger of the two confidence intervals. If interval 3 had been contained entirely within interval 1, there would have been a full overlap, yielding a consistency score of 1. Interval 3 overlaps the range from -19 to -7 of interval 1. This is an overlap of 12 percentage points out of 26 (the length of confidence interval 1), yielding a consistency score of 12/26 = 0.46. The consistency score is computed using the length of the larger of two confidence intervals as the denominator, and the share of this confidence interval covered by an overlapping, but smaller, confidence interval as the numerator. If two confidence intervals do not overlap at all, the consistency score is 0. If they are identical, the consistency score is 1. If one confidence interval is entirely contained within another, the consistency score is also 1, since such an overlap does not indicate any difference between two estimates in terms of direction or magnitude, merely in terms of precision.

To illustrate these concepts, consider the four estimates of effect presented in Figure 1. These estimates are taken from one of the cases used to illustrate the criteria of causality (14, 15). A 95% confidence interval is shown for each estimate of effect. All estimates are statistically significant, thus indicating the presence of a real effect (criterion 1). The estimates are also fully consistent with respect to the direction of the effect. Effect size varies between 2.6 and 8.8. These effect sizes indicate a fairly strong effect (criterion 2). The consistency of the four estimates of effect was assessed by computing the consistency score for the six pair-wise comparisons that can be made between the four estimates (1 vs 2, 1 vs 3, etc). A total score of 3.18 was obtained. The maximum score is 6 (if all four estimates had been perfectly consistent). Hence, the overall consistency score is 0.53 (criterion 3).

Causal direction (criterion 4) can be determined in terms of three operational criteria: (1) The temporal order between variables: a cause comes before the effect in time, (2) A priori considerations: age can be causally related to accidents, but not the other way around, and (3) A reversal of effect when a treatment is removed or reduced in intensity. Unfortunately, as will be discussed later, none of these criteria is by itself sufficient to establish causal direction.

Control for confounding (criterion 5) can be evaluated by answering the following three questions: (1) What are the most important potentially confounding factors that could affect a study employing this particular study design? (2) Is it likely that these confounding factors actually did confound study findings? (3) Did the study adequately control for the potentially confounding factors? This three stage procedure can be justified as follows. It is not realistic to expect any observational road safety evaluation study to control for every conceivable potentially confounding factor. Any reasonable assessment therefore needs to list explicitly the factors that are known to be potentially confounding. However, a potentially confounding factor will not always confound. It is therefore necessary to assess the likelihood that a potentially confounding factor actually did influence study findings. As an example, confounding by regression-to-the-mean is highly likely if treated sites were selected because they had a bad safety record, but improbable if an entire population was treated, e.g. if a new law applying to a large jurisdiction is introduced. If a study did control for important confounding factors and if the effect attributed to the treatment did not vanish, the study will be regarded as satisfying the criterion of control for confounding factors.

Description and measurement of the causal mechanism generating an effect (criterion 6) can be defined operationally in terms of an explicit identification of, and measurement of
changes in the target risk factors of a safety treatment as well similar identification and measurement of risk factors that may represent potential behavioural adaptation to the treatment.

Findings of road safety evaluation studies that contradict well-established laws of physics or of human perception and information processing should be regarded as less likely to be causal than findings supportive of such laws (criterion 7). Relevant laws of physics include, for example, the relationship between speed and stopping distance, the relationship between friction and stopping distance, the relationship between speed, mass and kinetic energy, and the relationship between ambient luminance level and visibility. All these relationships have, all else equal, clear implications concerning, for example, the relationship between speed and the number and severity of accidents, the relationship between road surface friction and the number of accidents, the relationship between the compatibility of groups of road users in terms of mass and the probability of fatal or serious injury, and the relative accident rate in darkness compared to daylight.

If the dose of a safety treatment can be measured (criterion 8), a dose-response pattern is found if the size of the effect attributed to the treatment is larger the greater the dose of the treatment.

Finally, if a clearly defined target group for a safety treatment exists, an effect should be found only within the target group and not outside of it (criterion 9).

EXAMPLES OF USE OF THE CRITERIA

To show the use of these criteria, two studies will be used as examples. The studies have been selected because there is reason to believe that they will be assessed differently according to the criteria of causality. The first study is an evaluation of the effects on accidents of roadside technical inspections of heavy vehicles in Norway (16). The paper presenting the study contains a discussion of causality, employing most of the criteria discussed in this paper.

Table 2 presents the evaluation of this study by means of the criteria of causality. Criterion 1, existence of an effect, is not satisfied, as none of a total of six estimates of effect presented in the original study (in Figures 1-3 and Tables 1-3 of the original study) was statistically significant at conventional levels. The study also failed criterion 2, strength of effect, as none of the effects estimated, including the effects of confounding factors, were statistically significant. Criterion 3 was satisfied, as all estimates of effect were consistent with respect to direction and magnitude. As far as causal direction is concerned (criterion 4), the criterion was partly fulfilled, in that the direction of change in safety was consistent with the direction of change in the frequency of inspections in 23 out of 36 cases. The study controlled for four potentially important confounding factors (trend, new drivers, business cycle, regression-to-the-mean), but failed to control for a fifth (other safety measures introduced) (criterion 5). The causal mechanism generating effects was not evaluated (criterion 6). While findings, generously interpreted, are consistent with theoretical expectations, the basis of these expectations is not very well established (criterion 7). A dose-response pattern was found, but its form (linear) was implausible, in particular if extrapolated beyond the range of variation covered by the data used in the study (criterion 8). Specificity of effect (criterion 9) could in principle have been assessed if the necessary data had been available. Since they were not, the study failed this criterion. The overall assessment is that the study provides only a weak basis for claiming that changes in the rate of inspections were causally related to changes in safety. One cannot rule out a causal influence of the roadside inspections, but the evidence suggests that other factors may be at least equally important. On the whole, the study of the effects of roadside technical inspections was of rather low quality, as far as providing a basis for causal inference is concerned. Whether it is possible or meaningful to quantify study quality is discussed below.
The second example to be discussed concerns a study of the effects of speed control measures by Mountain, Hirst and Maher (14, 15). Three treatments were studied: (1) Speed cameras, (2) Speed humps, and (3) Horizontal road deflections (“chicanes”), all of which were intended to improve compliance with speed limits and reduce the mean speed of traffic. The study was a before-and-after study employing the empirical Bayes (EB) design. The evaluation of the study in terms of the criteria of causality is reported in Table 3.

The study fulfilled all the criteria referring to statistical association between treatment and effect (criteria 1-3). By controlling for regression-to-the-mean, the study provides support for concluding that the causal direction was from treatment to effect, and not the other way around, as would be the case if abnormally high numbers of accidents before treatment caused the treatment to be introduced (criterion 4). The study controls fairly well for confounding factors (regression-to-the-mean, trend, changes in traffic volume), but fails to control for accident migration (criterion 5). The chief causal mechanism in the study is changes in speed, and these changes are well documented (criterion 6) and are found to be related to the size of the effect on accidents, thus showing a dose-response pattern (criterion 8). Study findings are consistent with theoretical expectations as far as the direction of effect is concerned, but there is an element of inconsistency in that effects on fatal and serious injury accidents appear to be slightly smaller than effects on all injury accidents (criterion 7). One of the criteria, specificity of effect (criterion 9), was not applicable to this study. It is more likely that the effects attributed to the safety treatments in this study were of a causal nature than in the study evaluating roadside technical inspections of heavy vehicles. The study clearly fulfilled five of eight criteria that were applicable, and partly fulfilled the other three.

**DISCUSSION**

Are causal inferences defensible at all in non-experimental studies? This question has preoccupied scientists in many fields for at least the last 50 years. In studies of, for example, the causes of disease, the causes of crime, the effects of social deprivation on health or the quality of life, the scope for employing randomised controlled trials is highly limited. However, there is a demand for knowledge in these fields, even if this knowledge has to be gained from observational studies.

In discussing the widely known criteria of causality that have been developed in epidemiology, Rothman and Greenland (4) stress the limitations of these criteria and remind readers that some of them may be inconclusive or difficult to apply. This scepticism is well-placed with respect to the criteria discussed in this paper as well. Consider, for example, the first criterion, stating that there should be a statistical relationship between cause (treatment) and effect. It is not difficult to think of cases in which one could reasonably argue that a causal influence was present, even if it did not manifest itself in the form of statistically detectable changes in the number of accidents. An interesting case of this is a study by Amundsen (17), concerning the effects on safety of bright road surfaces. Bright road surfaces are intended is to improve reflection, thus increasing sight distances (in particular in dark) and improving the contrast between the road and its surroundings. The study found that bright road surfaces did influence the target risk factors in the intended direction: sight distances increased by a mean of 16%, and cars moved about 2% away from the edge-line of the road. However, drivers adapted their behaviour by increasing speed by about 3%. The net result was that the number of accidents was almost unchanged; there was an increase of 1%. In this case, it is reasonable to say that the treatment did have an effect, but the causal mechanism generating the effects of the treatment consisted of several partial effects interacting in a way that cancelled the effect on accidents.
It is commonly believed that causation does not operate backwards in time; hence if the temporal order between variables can be determined, causal inferences are supported. Yet, not even this criterion is watertight. Suppose that at time T it is announced that from time T + 1, a graduated driver licence system will be introduced. New drivers will then have to go through more training, and accept a restricted licence, before getting full driving privileges. It would not be surprising if, during the interval between time T and time T + 1, there was a rush of people obtaining licences. Thus, in a sense, the measure would have an effect before it formally took effect. Expectations of events to come can generate effects before the events actually occur.

Control for confounding variables is, arguably, the Achilles heel of observational road safety evaluation studies. The control for confounding variables is always imperfect in observational studies. In principle, it is possible to control for several potentially confounding variables at the same time by developing multivariate accident prediction models. In a critical analysis of such models, Hauer (8) questions whether it is possible to draw causal inferences based on them. He notes that the findings of different models developed to answer the same question (e.g. how do automatic gates influence the safety of highway-railroad grade crossings?) are often inconsistent. There is no consensus on how to develop the “right” model. Models differ in terms of the variables included, the specification of the relationship between these variables, the functional forms assumed, the assumptions made regarding the residual terms, and so on.

While the criteria of causality proposed in this paper can be applied to evaluate a multivariate model, it seems clear that additional criteria are needed. A problem likely to affect most multivariate models developed to evaluate the effects of road safety treatments is endogeneity. This is the problem caused by bias-by-selection. Safety treatments tend to be introduced at high-risk locations. Even if the treatments are effective, the expected number of accidents after treatment may remain higher at the treated locations than at untreated locations. A paper by Kim and Washington (18) shows the errors that may result if researchers are unaware of endogeneity and do not control for it in accident prediction models. Using left-turn lanes as an example, Kim and Washington show that with no control for endogeneity, left turn lanes are estimated to be associated with an increased number of accidents. With control for endogeneity, the coefficient changed sign and left-turn lanes are estimated to be associated with fewer accidents. Another instructive example of the importance of recognising and controlling for endogeneity is given by Taylor et al (19).

An intricate issue arising when evaluating multivariate models is that some of the statistical relationships estimated in such models can reasonably be thought of as causal, while other relationships are not causal. The criteria proposed in this paper may help identify those relationships that are causal and those that are not. Yet, this might not be easy. Does traffic cause accidents? Yes, it does, by being a necessary condition for accidents. What, then, is the causal mechanism by which traffic causes accidents? It is a complex mechanism, in which laws of physics and laws of human perception and information processing contribute importantly. Few, if any, studies have uncovered this causal mechanism in detail. In fact, most multivariate models fail to address causal mechanisms at all, leaving them as a black box. Thus, even for such a basic and widely accepted relationship it would be difficult to find studies that have assessed the relationship between traffic and accidents in a sufficiently rigorous manner to fulfil all or most of the criteria of causality.

When applied to a single study, the criteria of causality may rarely be conclusive since it will almost never be the case that an observational study meets all the criteria perfectly. When applied to a set of studies, the criteria could be more helpful by directing attention to patterns in study findings that are only apparent when multiple studies are considered. For example, a dose-response pattern may sometimes be observed when the findings of several studies are compared, but not within a single study.
It will typically be the case, as shown in the two examples used in this paper, that a study fulfils some of the criteria very well, some of them halfway, and some of them not at all. What should one conclude in such cases? Is it possible to develop an overall score indicating study quality based on the criteria?

Developing an overall score is no problem. It can be done simply by assigning numbers to the criteria and converting the verbal assessment to a numerical score. The trouble, as pointed out by Greenland (20), is that any such numerical score will be arbitrary. Different researchers may assign different weights to the criteria, resulting in different scores for the same study. Trying to get a consensus on a numerical scoring system is difficult. Researchers may agree that, for example, control for confounding is very important. But how important is it? Should it carry 60% of the sum of weights given to the criteria or 80%? It is hard to give a very good justification for choosing one or the other.

Even if a widely accepted scoring system could be developed, there might still be room for disagreement. Does a study getting 60% of the maximum score support a causal inference, or does it not? Some researchers may be reluctant to infer causality unless the score in favour of doing so is at least 80%; others may be willing to do so if the score favouring it is only 60%.

Despite all these problems, the criteria of causality discussed in this paper may be useful. At the very least, they can be used as a checklist for a systematic discussion of the findings of a study. That discussion might not always be conclusive, but it would at least be informative. It would also point to elements of study design or analysis that should be improved in future studies to provide a better basis for inferring causal relationships. Observational studies, no matter how well executed, can never be as conclusive with respect to causality as a well-executed randomised controlled trial. Some observational studies will, however, provide stronger evidence in favour of causality than others; a systematic assessment by means of explicit criteria can help identify those studies.

CONCLUSIONS

The main conclusions emerging from this paper are as follows:

1. It is possible to define in theoretical and operational terms criteria of causality for observational road safety studies, based on similar criteria that have been developed and applied in epidemiology.

2. The criteria of causality were applied to two test cases. These cases indicate that the criteria are able to discriminate between studies providing a relatively firm basis for causal inference and studies providing a weak basis for causal inference.

3. Additional criteria are likely to be needed for studies employing advanced statistical modelling.

4. It is likely that the criteria of causality will not always be conclusive, either because their application to a specific study is not straightforward or because the study will be found to fulfil only some of the criteria.

REFERENCES


FIGURES AND TABLES

**Figure 1**: Assessing the statistical association between treatment and effect – consistency of estimates of effect

**Table 1**: Criteria of causality for observational road safety evaluation studies

**Table 2**: Application of criteria of causality for observational road safety evaluation studies to a study by Elvik (16)

**Table 3**: Application of criteria of causality for observational road safety evaluation studies to a study by Mountain et al. (14, 15)

**Figure 1**: Assessing the statistical association between treatment and effect – consistency of estimates of effect

![Graph showing percentage change in the number of injury accidents for different effect sizes.](image)

- Estimates 1 and 2 are fully consistent (smaller confidence entirely within larger)
- Estimates 3 and 4 are partly consistent (confidence intervals partly overlap)

*Effect size = 22/4 = 5.5, Effect size = 29/11 = 2.6, Effect size = 44/5 = 8.8, Effect size = 37/6 = 6.2*
### TABLE 1: Criteria of causality for observational road safety evaluation studies

<table>
<thead>
<tr>
<th>Criterion of causality</th>
<th>Theoretical definition</th>
<th>Operational definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Statistical association</td>
<td>There should be a statistical association between cause and effect</td>
<td>A statistically significant change in variables measuring safety associated with safety treatment</td>
</tr>
<tr>
<td>2. Strength of association</td>
<td>A strong association is more likely to be causal than a weak association</td>
<td>Treatment effect stated in terms of effect size compared to effect sizes for other variables present in the data</td>
</tr>
<tr>
<td>3. Consistency of association</td>
<td>A consistent association is more likely to causal than an inconsistent association</td>
<td>The consistency in direction and size of effect attributed to safety treatment across subsets of the data or different model specifications, assessed by means of a consistency score (see text)</td>
</tr>
<tr>
<td>4. Clear causal direction</td>
<td>It should be clear which of two variables is the cause and which is the effect</td>
<td>The temporal order between variables; a priori considerations; reversal of effect when treatment is removed</td>
</tr>
<tr>
<td>5. Control for confounders</td>
<td>The association between cause and effect should not vanish when confounding variables are controlled for</td>
<td>The identification of potentially confounding variables; invariance of the effect attributed to treatment with respect to potentially confounding variables controlled for; completeness of the control for confounding variables</td>
</tr>
<tr>
<td>6. Causal mechanism</td>
<td>The mechanism generating an effect should be identified and measured</td>
<td>Changes in target risk factors influenced by a road safety treatment and changes in risk factors representing behavioural adaptation to the treatment</td>
</tr>
<tr>
<td>7. Theoretical explanation</td>
<td>A plausible theoretical explanation of the findings of a study should be given</td>
<td>Findings should not contradict well established laws of physics or laws of human perception and information processing</td>
</tr>
<tr>
<td>8. Dose-response pattern</td>
<td>Treatments administered in large dose should have larger effects than treatments administered in small doses</td>
<td>Treatments that are intense or have large effects on target risk factors should be associated with larger changes in safety than less intense treatments or treatments with small effects on target risk factors</td>
</tr>
<tr>
<td>9. Specificity of effect</td>
<td>Effects of a cause operating only in a certain clearly defined group should only be found within that group</td>
<td>An effect of safety treatments targeted at clearly defined groups should only be found in those groups and not in other groups</td>
</tr>
</tbody>
</table>
**TABLE 2: Application of criteria of causality for observational road safety evaluation studies to a study by Elvik (16)**

<table>
<thead>
<tr>
<th>Criterion of causality</th>
<th>Application of the criterion</th>
<th>Conclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Statistical association</td>
<td>The statistical association between treatment and effect is shown in three figures and three tables of the original paper (16). There is an association between treatment and effect in all six cases shown, but it is in no case statistically significant.</td>
<td>Criterion not fulfilled.</td>
</tr>
<tr>
<td>2. Strength of association</td>
<td>The change in the number of accidents associated with a 100% change in the annual number of inspections per vehicle is small and not statistically significant. Effects of other variables were not consistently stronger than the effects of inspections.</td>
<td>Criterion not fulfilled.</td>
</tr>
<tr>
<td>3. Consistency of association</td>
<td>The direction and magnitude of the effect attributed to roadside inspections is the same for all six tests presented.</td>
<td>Criterion fulfilled.</td>
</tr>
<tr>
<td>4. Clear causal direction</td>
<td>Inspections increased in 7 out of 12 cases; decreased in 5 out of 12 cases. In figure 1 of the original paper, the direction the association between changes in the number of inspections and changes in accidents was consistent with a clear causal direction (i.e. more inspections = lower accident rate, and less inspections = higher accident rate) in 9 out 12 cases. By the same token, consistency was 7 out of 12 cases in Figure 2, and 7 out of 12 cases in Figure 3 of the original paper. In total, direction was consistent in 23 out of 36 cases.</td>
<td>Criterion partly fulfilled.</td>
</tr>
<tr>
<td>5. Control for confounders</td>
<td>The study controlled for (1) long-term trends, (2) recruitment of new drivers, and (3) economic growth by means of a regression model. Also (4) regression-to-the-mean (a high accident rate in year N leading to more inspections in year N + 1) was checked for by lagged correlations. The effects attributed to inspections were weaker when 3 potentially confounding variables were controlled for than when 0 potentially confounding variables were controlled for. Two potentially confounding variables not controlled for by the study were (1) changes in enforcement of service and rest hours (increased towards the end of the period) and (2) the introduction of a top speed governor.</td>
<td>Criterion partly fulfilled.</td>
</tr>
<tr>
<td>6. Causal mechanism</td>
<td>The study did not uncover the causal mechanism generating the effect. There were no data on technical condition and how it was affected by inspections. There were no data on potential behavioural adaptation by drivers.</td>
<td>Criterion not fulfilled.</td>
</tr>
<tr>
<td>7. Theoretical explanation</td>
<td>No well-established theory exists to explain study findings. Findings are, however, consistent with studies that have found that heavy vehicles with technical defects are more often involved in accidents than heavy vehicles in good technical condition.</td>
<td>Criterion partly fulfilled.</td>
</tr>
<tr>
<td>8. Dose-response pattern</td>
<td>The dose of inspections varied from 0.39 per vehicle per year to 1.88 per vehicle per year. A dose-response pattern was found, but it was not statistically significant and the form of the relationship (linear) is implausible when extrapolated.</td>
<td>Criterion partly fulfilled</td>
</tr>
<tr>
<td>9. Specificity of effect</td>
<td>Specificity in the present case would obtain if an effect was found only in defective vehicles and not in vehicles not found to be defective. There were no data to assess the specificity of an effect.</td>
<td>Criterion not fulfilled.</td>
</tr>
<tr>
<td>Criterion of causality</td>
<td>Application of the criterion</td>
<td>Conclusion</td>
</tr>
<tr>
<td>------------------------</td>
<td>------------------------------</td>
<td>------------</td>
</tr>
<tr>
<td>1. Statistical association</td>
<td>Table 3 of the original paper reports 8 tests of the effects of the safety treatments evaluated. The overall effect attributed to the treatments was statistically significant in 6 of the 8 cases.</td>
<td>Criterion fulfilled.</td>
</tr>
<tr>
<td>2. Strength of association</td>
<td>Overall effects attributed to safety treatments varied between 11% accident reduction and 44% accident reduction. The effects of two confounding variables, long-term trends and regression-to-the-mean, were estimated to be considerably weaker.</td>
<td>Criterion fulfilled.</td>
</tr>
<tr>
<td>3. Consistency of association</td>
<td>The direction of the overall effects attributed to safety treatments were consistently in the same direction. The magnitude of the effects differed significantly between types of treatment. Sources of variation in magnitude of effect were identified.</td>
<td>Criterion fulfilled.</td>
</tr>
<tr>
<td>4. Clear causal direction</td>
<td>The study employed a before-and-after design controlling for regression-to-the-mean. The nature of the safety schemes evaluated makes it unlikely that backwards causation could be present (see text for a discussion of what backwards causation means).</td>
<td>Criterion fulfilled.</td>
</tr>
<tr>
<td>5. Control for confounders</td>
<td>The study controlled for long-term trends and regression-to-the-mean. Changes in traffic volume were also accounted for, but were treated as an endogenous variable, i.e. as one of the effects of the safety treatments. Potentially confounding variables not controlled for include (1) changes over time in accident reporting (in practice almost impossible to control for), (2) the use of other safety treatments in addition to those evaluated, and (3) the possible presence of accident migration (due to diversion of traffic to alternative routes).</td>
<td>Criterion partly fulfilled.</td>
</tr>
<tr>
<td>6. Causal mechanism</td>
<td>The two main causal mechanisms through which the treatments influenced the number and severity of accidents were changes in speed and changes in traffic volume. Both were identified and measured. Changes in speed appeared to be the dominant mechanism, accounting for a substantially larger share of the net change in accidents attributed to the treatments than changes in traffic volume. Behavioural adaptation might take the form of, for example, displaced traffic choosing alternative routes without the speed control devices. Such behavioural adaptation could lead to accident migration. The study did not address this issue.</td>
<td>Criterion fulfilled.</td>
</tr>
<tr>
<td>7. Theoretical explanation</td>
<td>Speed was reduced. The number of injury accidents was also reduced. This is consistent with the laws of physics. However, the percentage reduction of the number of fatal and serious injury accidents was smaller than the reduction of all injury accidents, which is inconsistent with the expectation that injury severity is reduced when speed is reduced.</td>
<td>Criterion partly fulfilled.</td>
</tr>
<tr>
<td>8. Dose-response pattern</td>
<td>The study investigated the presence of a dose-response pattern by investigating the relationship between changes in speed and changes in the number of accidents. A clear dose-response pattern was found for two of the three safety treatments, but not for the third.</td>
<td>Criterion partly fulfilled.</td>
</tr>
<tr>
<td>9. Specificity of effect</td>
<td>Safety treatments influencing speed will have an effect on all accidents. Hence, the specificity criterion is inapplicable for this study.</td>
<td>Not applicable.</td>
</tr>
</tbody>
</table>