Presented at the Road Safety Research, Policing and Education Conference,
held in Perth, Western Australia, 14th-16th November 2004

(Session on Research Methodologies)

Control groups, randomisation, double-blinding, meta-analysis: Can road safety research learn from evidence-based medicine and social welfare?

T. P. Hutchinson
Centre for Automotive Safety Research, University of Adelaide, South Australia 5005
(paul@casr.adelaide.edu.au; +61 8 8303 5997; fax +61 8 8232 4995)

Abstract

The practice of medicine, these days, is supposed to be evidence-based. Important features of this are the use of randomised experimentation to compare treatment and control interventions, and the systematic reviewing of such research using meta-analysis. The idea has spread to social welfare, criminology, and education. The present paper considers whether evaluation of road safety interventions should use similar methods, or whether evidence-based everything is a fad that we can safely ignore, impracticable in road safety. A real-world road safety intervention is typically a much more complex process than, say, a comparison of aspirin with placebo. However, some of the experiments in social welfare, criminology, and education must have been just as difficult as those in road safety. There is much in the statistical, medical, and social welfare literatures on the desirability of a control group, randomisation, and double-blinding, and the dangers of merely making a before-after comparison, of allocating experimental units to groups in a non-random manner, and of allowing either the experimental units or the researchers to know which is in which group. However, there are many issues in transport safety for which the disadvantages of a rigorous methodology will outweigh the advantages. Nevertheless, for some issues, randomisation (and other precautions) ought to receive more consideration than at present. If that happens, planning the monitoring of a road safety intervention will take longer and need more resources.

1. Introduction

The phrase "evidence-based", in this paper, has a specific meaning. It refers to research of high quality methodologically --- notably, that conducted using randomised experimentation --- and the systematic reviewing of such research using meta-analysis. It is not merely a general expression of hope that there will be some input to policy of objective data. In medicine, the phrase came to prominence in the early 1990's. The ideas are now also influential in social welfare fields, with slight change of emphasis. In medicine, "evidence-based" implies other things also, such as double-blinding (concealment of allocation), the practitioner keeping up-to-date with research, the practitioner considering the individual patient in the light of that research, and a degree of emphasis on both the value and limitations of quantitative methods in (for example) diagnosis and the taking of treatment decisions. In social welfare, these additional elements tend to be less important: double-blinding is largely impracticable, and there usually is noone corresponding to a medical practitioner. But whether in medicine or other fields, randomised experimentation and systematic reviewing are the most distinctive features.

Another catchcry, especially in social welfare contexts, is "what works and what doesn't work". The intent of the phrase seems to be to emphasise the practical: decide on a dependent variable (outcome) that is objectively measured, try to avoid having many exclusion criteria in defining the population of interest (in order that the results be of wide relevance), and recruit a large sample size (in order that the results be reasonably precise). Because of the difficulties of researching the impact on humans of other humans' implementation of policies, there is a preference for simplicity of experimental design.

The aim of the present paper is to consider the practicability of higher methodological standards in research into interventions in the transport and transport safety fields:

> "If road safety policy is to be based on the best available research evidence, there must be a greater understanding that study design has an important bearing on the validity of research findings" (Cochrane Injuries Group Driver Education Reviewers, 2001, p.232).

Should we wholeheartedly advocate randomisation and meta-analysis, and aim for evidence-based policy? Or can we dismiss evidence-based everything as a fad that we can safely ignore, impracticable in road safety? The present paper discusses these questions. The answers given are admittedly subjective.

2. How research should be conducted

The purpose of this Section is to list and justify some important features of how research (research that evaluates interventions, that is) should be conducted.

- Specify the dependent variable (the outcome) of interest.
- Identify the unit to which the intervention is being directed. For example, the units might be people (or in a transport example, intersections)
- Define all the units of interest, and randomly assign each unit (or a random sample) to either the treatment group or the control group.
- Measure the present condition of each unit.

- Apply the intervention to the units in the treatment group (and do nothing to those in the control group).
- Conceal, both from the people participating and from the researchers who are evaluating the outcomes, which units are in which of the groups. That is, conduct the research in a double-blind fashion.
- Measure the condition of each unit again.
- The change of each unit is now known. The changes in the treatment group can now be compared with the changes in the control group.

Randomisation. If any process other than randomisation is used to allocate experimental units to treatment and control groups, it cannot be known to be unbiased. There are many examples in the medical and social science literatures that show how biases occur when randomisation is not attempted or is attempted but fails. (The experimental design may involve randomly allocating n units to one group and n to the other, and the present paper was written with this in mind. But other designs are also used: in particular, there may be stratification according to one or more background factors, with randomisation taking place within each stratum; taken to an extreme, this becomes a matched-pairs design.) Of course, there remains the issue of chance: there will be random differences between the groups in respect of background factors that may affect the dependent variable --- but these are sure to be small when the sample size is large. This is true both for background factors that we might know about and might be able to take account of in ways other than randomisation, and for ones that we did not imagine might be relevant, or are unable to measure. An alternative strategy is not to randomise but to adjust for differences between the two groups. Glonek (2001) is sceptical, as many others have been, about this, and notes that the limited data available in real studies may make it impossible to choose between different models that imply incompatible conclusions; and, he continues, "In any observational study or non-randomised experiment, there may be confounding variables that have a large impact on the apparent treatment differences but have not been recorded. Such variables clearly cannot be adjusted for, so it can only be concluded that any conclusions drawn from observational studies or non-randomised experiments must be to some extent speculative".

Regression towards the mean. One particular danger if randomisation is not employed is regression towards the mean. That is, there may be a tendency for units that have the problem to a severe degree (whether they are people with a high blood pressure, schools with a high rate of delinquency, or road sections with a high accident rate) to be allocated to the treatment rather than the control group. Part of the reason these units seem to be problem cases is mere chance. After treatment, there appears to be an improvement, but this is because they were not really problem cases to begin with. It has been appreciated for many years (e.g., Tamburri et al., 1968) that it is inappropriate to use the number of accidents in previous years as a basis for allocation of sites to treatment or control groups: for some of the sites that had a lot of accidents in previous years, this was merely a matter of chance, and they would return to a more normal accident level whether treated or not. (Outside of a research context, it may be sensible to use a high number of accidents as a criterion for treating a site, as in many "black spot" programmes.)

Sample size. Mention should also be made of sample size. There is only limited value in doing research with a sample size too small to answer the question with

reasonable accuracy. When planning the research, if the aim is to estimate the size of an effect, the sample needs to be large enough to lead to a standard error of the estimate that is small enough for whatever the specific purpose of the study is. If the intention is to carry out a statistical test, the standard error calculations can be taken further and expressed as the power of the test. (Small studies can have value when put together in a meta-analysis. But if it is decided to conduct an experiment on the off-chance it will be fairly conclusive itself, but more realistically will only provide a little evidence about the research question, the limitations should be known by the experimenters and their ethics committees.) It may be added that sometimes there is criticism of research projects for using too large a sample size, as this is wasteful of resources, and delays the benefits from answering the question.

3. Evidence-based medical, social welfare, and road safety research

Evidence-based medicine and the Cochrane Collaboration. In recent years there has been a trend to higher methodological standards in medical research, because of the biases that can easily creep into the comparisons that are of central interest in a research project. "Evidence-based medicine" has several elements, but the most distinctive and contentious are systematic reviewing of previous research on a topic, giving much greater weight to studies that adhere to high methodological standards than to those of lower standards, and attempting to conduct new research to high standards rather than lower, especially the preferring of randomised trials over observational studies.

- *Systematic review.* A really extensive search for relevant previous research is conducted, making full use of computerised databases and software to interrogate them. The methodological quality of the papers and reports found is evaluated. For those of sufficiently high quality, the results found are summarised by a meta-analysis (i.e., by some appropriate quantitative method, such as taking the average), rather than a review in narrative style.
- *Randomisation.* High methodological standards are advocated in the conduct of research, including obtaining a large enough sample size to answer the question of interest, randomised allocation of the experimental units (these may or may not be individual patients) to treatment or control groups, and blinding of both the units and the experimenter to the allocation.

The Cochrane Collaboration (http://www.cochrane.org/index0.htm) is part of this trend. This "aims to help people make well informed decisions about health care by preparing, maintaining, and ensuring the accessibility of systematic reviews of the effects of health care interventions" (Chalmers et al., 1998).

Evidence-based social welfare research and the Campbell Collaboration. An editorial in the Journal of the Royal Statistical Society by Fitz-Gibbon (2004) advocates more real-world social experimentation with randomised assignment, and more systematic reviews of the Cochrane Collaboration type. It notes that the Campbell Collaboration (http://www.campbellcollaboration.org) has been established to promote systematic reviews of research in the social sciences. According to Fitz-Gibbon, "Many examples illustrate that guessing and good intentions are not a basis for effective action.... we must check our theories and hypotheses". Humility is a recurring theme in writings by advocates of randomised experiments and meta-analysis (e.g., Chalmers, 2003). They emphasise that there are many instances in

medicine, criminology, education, social welfare, and so on where noone --- not the experts, the administrators, the politicians, the pressure groups --- really knows which is the best treatment. Plausibility, recommendation by experts, and empirical support from studies of low methodological quality are not enough, and can be deadly if they deter the rigorous evaluation of an intervention. Randomised experimentation is not new: Farrington (2003) regards so early a period as 1965-1975 as having been the golden age of British randomised experiments on crime. Here are two examples of the contrast that can exist between popular beliefs and systematic reviews.

- "Scared straight" programmes (in the U.S.A.) involve juvenile delinquents visiting adult prisons, getting scared by what they experience, and then (supposedly) being deterred from a life of crime. However, a systematic review by Petrosino et al. (2003) has concluded that such programmes tend to increase delinquency. Such programmes continue to be used, supported by evidence of an indirect nature (namely, what prisoners and the participants say about them). Petrosino et al. (2003) note that shock value interventions are also tried in other fields, including road safety, and they wonder if the results are disappointing or even toxic, as with scared straight programmes.
- And in our own field, a systematic review by Roberts et al. (2004) found that driver education leads to earlier licensing, and that there is no evidence that it reduces road crash involvement.

There are useful collections of papers on the evidence-based movement in social welfare edited by Mosteller and Boruch (2002) and Sherman (2003).

Evidence-based road safety research. Morrison et al. (2003) identified 28 systematic reviews on improving population health through transport interventions. These were classified as health promotion (14 studies), engineering (5 studies), environmental (3 studies), or legislative (6 studies). Morrison et al. found that the following can improve health: health promotion, traffic calming, some legislation; but some interventions, such as driver improvement and education courses, may be harmful. Hutchinson and Meier (2004) reviewed transport and transport safety research that has been influenced by the evidence-based methods. They listed 17 Cochrane reviews either completed or in progress, and 17 traffic engineering papers showing the influence of evidence-based medicine. The following points are largely based on their paper.

- Transport and transport safety research is being openly criticised by some people for not making greater use of randomised experimentation.
- On some topics, especially those close to medicine, where randomisation of individual people is possible, it is quite common. See especially Supplement 1, Volume 16 (1999), of the American Journal of Preventive Medicine, and Supplement 4, Volume 21 (2001) of the same journal.
- Randomisation and treatment of units other than individual people is rare, but does sometimes happen. Some experiments pre-dated the phrase "evidence-based", e.g., that of Rausch et al. (1982), who randomised taxi-cabs. Retting et al. (2002) randomised intersections. Sometimes groups of people (e.g., a school) are randomised when the treatment is one of individuals.
- Meta-analysis is common, but often is not fully followed through, in the sense that studies of quite low methodological quality are retained in the sample. This practice may be justified as making the best use of what evidence is actually available, but it does not sit very easily with the demand for conclusions based upon high quality evidence.

4. Some problems with randomisation and meta-analysis in road safety

Randomised experimentation: principled criticisms. When it moves from drug trials to public health and social welfare, randomised experimentation becomes open to a number of principled criticisms.

- *Standardisation of the intervention.* Road safety interventions are not as standardised as giving a specific dose of a specific drug. Variations may be important. (However, a paper by Hawe et al., 2004, argues that it may be possible to define an implementation of an intervention that is standardised by function rather than by composition or form.)
- *Interaction of the intervention with other factors.* Even if the intervention can be closely specified so that it is the same everywhere, its effect may be affected by social, economic, and geographical circumstances. Indeed, it may be affected by what the baseline level of the dependent variable (e.g., accident rate) is. Consequently, the relevance of the experiment to different circumstances may be questionable.
- *Chance differences in background variables after randomisation.* If the chance differences are appreciable in magnitude, it may be plausible that they are affecting the dependent variable. Their occurrence is especially likely when the total sample of units available is small, which is frequent in road safety research. The arguments for randomised experimentation are a lot more persuasive when the sample size is large than when it is small.
- *Statistical testing.* Evidence-based medicine makes much use of the results of statistical tests, and thus its credibility is impacted by the criticisms that are made of these. Level of significance (p-value), and conclusion drawn, can only be taken at face value if the hypothesis is specified in advance of seeing the data, yet many details of the testing are usually decided only in course of processing the data.

The opinion of the present writer is that in the context of transport and transport safety research these criticisms are intellectually respectable; they may or may not constitute good arguments, depending upon the particular issue being considered; it cannot simply be said that the evidence-based approach is the only correct one.

Randomised experimentation: practicability. The nature and seriousness of the practical objections to randomised experiments in road safety vary greatly from one issue to another. They seem valid, yet could be overcome, albeit with greater planning and organisation than is typical at present. Three references from the social welfare literature that are helpful concerning practical details of experimentation are Watson et al. (2004), Oakley et al. (2003), and Peterson et al. (2000).

Meta-analysis. Among the criticisms that have been made of meta-analysis, the following seem particularly relevant in road safety.

- In practice, meta-analysis seems to go far beyond only combining studies that were very similar in methodology in order to reduce random variation. It is common for road safety studies of one issue to be quite different from one another. Thus it is far from clear that they really should be averaged, as a meta-analysis usually does. A vivid illustration of the problem is that Thomas Edison failed many times in attempting to invent a light bulb, before eventually succeeding: the important thing is that the final attempt succeeded, not that the average was failure (Sherman, 2003, Preface). Similarly, in a meta-

analysis of an intervention whose implementations are slightly different from each other, it may be than the one instance of success is the signpost to future development, not the average outcome of failure.

- The medical literature emphasises high quality evidence, but also the best available evidence. There is ambiguity here, which is unfortunate when considering road safety. In this field, the number of research studies on any particular issue that are of high quality (by the criteria usual in medicine) is close to zero, so it is important to make distinctions at the lower quality end of the scale. However, this is not easily done. (a) Concerning non-randomised before-after treatment-control comparisons, the evidence from these is usually considered to have some value if the control group really is an appropriate one, but to be almost valueless if that is not the case. Unfortunately, scales of research quality do not give assistance in judging the appropriateness of the control group. (b) When there is no other evidence available, we might look at naive (i.e., without a control group) before-after comparisons. There is no general agreement about whether these are worth something or nothing.

5. Contrasting positions about evidence-based road safety

What view might we take about the relevance to road safety of evidence-based medicine and social welfare? Let's get some fixed points by describing the extremes.

- We might take to heart all the literature informing us of the desirability of a control group, randomisation, and double-blinding, and the dangers of merely making a before-after comparison, of allocating experimental units to groups in a non-random manner, and of allowing either the experimental units or the researchers to know which is in which group. We would conclude that the poor methodological standard of the vast majority of research implies that researchers have been wasting their time. And we would campaign for evidence-based policy and practice.
- Or we might give credence to several important objections of principle and of practicality to evidence-based everything. We might consider that in transport and transport safety, there are good reasons for departing from textbook recommendations about how research should be conducted. We might observe that just because biases could potentially creep in, this does not mean that they actually have crept in, and we might express the opinion that there is no evidence that wrong conclusions occur often.

A position somewhere between these extremes seems appropriate. Many issues in transport and transport safety are likely to raise insurmountable problems for randomisation. Nevertheless, the claims that are made for randomised experimentation are largely valid (in my opinion), and it ought to be more seriously considered than at present. If the benefits were more widely known, or the disadvantages of other methodologies were taken more seriously, then more opportunities for utilising randomisation, and a greater willingness to overcome the difficulties, might possibly emerge. If so, planning the research (listing alternative methodologies, considering their merits, conducting pilot studies, identifying the problems, and judging whether the problems can be overcome) may need to take longer and consume more resources than it typically does at present. Moreover,

greater desire to use randomised experimentation may necessitate closer cooperation between several jurisdictions on a given research question.

It seems that there is wide acceptance of the benefits of a double comparison, that is, the before-after change in a treatment group and in a control group. If this is both practicable and desirable, why is the extra step to randomised allocation so rarely taken? If it is that politicians or the press or the public would find this objectionable, this might be overcome by education that draws on the medical analogy: when we do not know if a drug works or not, we randomise patients to a drug or no-drug condition, and when we do not know if a road surface treatment works or not, we should similarly randomise lengths of road to treatment or no-treatment conditions. If, on the other hand, the rarity of randomisation is because of the extra formal advance planning of research that is involved, then the solution lies within the research community. It may be that more effort should be put into the design and planning of research generally, and into coordination between different places. When evaluation is wanted, it needs to be taken seriously, it is no good trying to do the job on the cheap. Oakley et al. (2003) have noted the need for adequate resourcing of development time in setting up randomised experiments. Researchers and those funding research need to consider how greater planning of research projects and greater coordination of efforts in different places would be paid for. And this is not only a matter of finding the dollars: there may be a difference in the mentality and attitudes of a researcher and a research planner.

Are evidence-based methods a threat or an opportunity in transport and transport safety research? In my view, both.

- The threat is that much research may be dismissed as valueless because it was not an experiment with randomised treatment and control groups. However, randomised experiments have their own set of problems as well as advantages. These need to be weighed up, as do the problems and advantages of any methodology. By all means, criticise what should be criticised; but do not imply that there always has been or is now an alternative methodology (randomised experimentation) that is free from disadvantages.
- The opportunity is that for some questions, a randomised experiment is indeed the methodology of preference, yet has been under-utilised in transport and transport safety. The current trend towards evidence-based methods in social welfare research may help to overcome the reluctance to utilise these methods. (For example, those commissioning research may become accustomed to more expensive research, or irrational prejudices against randomisation may weaken.)
- This opportunity could be expressed more forcefully: according to Hauer (1988), the basic data relevant to many common decisions in traffic engineering relating to safety do not exist, or, when they do exist, are ignored. If things really are this bad, we should welcome the evidence-based everything movement as much for its inspirational and complacency-busting aspects as for its methodological merits.
- Research cannot be of high quality if the sample size is too small to answer the question. A demand for high quality research may lead to more appropriate choice of sample size than at present.

6. Present and future

When submitting abstracts to this Conference, authors were asked also to provide a "rationale" for their paper. What I said was:

> The use of accident statistics and measures of behaviour to determine what effect a real-world intervention has had is one of the major methodologies in road safety research. Many delegates will be conscious that the research they do is far less rigorous than clinical trials of new drugs, for example. Very likely, they have been asking themselves how guilty should they feel, and whether they can do better. This paper will tell them.

So: how guilty should you, the expert road safety researcher, feel? At present, there is not much reason for guilt. The diffusion of evidence-based methods from medicine into other fields has been cautious, and for good reason, so road safety has not fallen behind. And you have read this paper, which is a way of checking that you have not fallen behind. As to the future, though, there will be reason for guilt if attention is not paid to the success that evidence-based methods are beginning to have in various social welfare fields, and serious consideration given to whether they may be the right choice for particular projects in road safety research.

Acknowledgements

The Centre for Automotive Safety Research receives core funding from both DTUP and the South Australian Motor Accident Commission. The views expressed in this report are those of the author and do not necessarily represent those of the University of Adelaide or the sponsoring organisations.

References

Chalmers, I. (2003). Trying to do more good than harm in policy and practice: The role of rigorous, transparent, up-to-date evaluations. *Annals of the American Academy of Political and Social Science*, 589, 22-40.

Chalmers, I., Sackett, D., and Silagy, C. (1998). Cochrane Collaboration. In *Encyclopedia of Biostatistics. Volume 2*, pp. 764-767. Chichester: Wiley.

Cochrane Injuries Group Driver Education Reviewers (2001). Evidence based road safety: The Driving Standards Agency's schools programme. *The Lancet*, 358, 230-232. (This is reproduced in *Proceedings of the 3rd International Interdisciplinary Evidence-Based Policies and Indicator Systems Conference*, 2001, pp. 8-13. Durham: Curriculum, Evaluation and Management Centre, University of Durham.)

Farrington, D. P. (2003). British randomized experiments on crime and justice. *Annals of the American Academy of Political and Social Science*, 589, 150-167.

Fitz-Gibbon, C. (2004). Editorial: The need for randomized trials in social research. *Journal of the Royal Statistical Society, Series A*, 167, 1-4.

Glonek, G. (2001). On the strengths and weaknesses of experimental and observational methodologies in road safety research. Presented at the NRMA Insurance National Speed and Road Safety Conference, held in Adelaide.

Hauer, E. (1988). A case for science-based road safety design and management. In Stammer, R. E. (Editor), *Highway Safety: At the Crossroads*, pp. 241-267. New York: American Society of Civil Engineers.

Hawe, P., Shiell, A., and Riley, T. (2004). Complex interventions: How "out of control" can a randomised controlled trial be? *British Medical Journal*, 328, 1561-1563.

Hutchinson, T. P., and Meier, A. J. (2004). Evidence-based road safety policy? Evidence-based transport policy? A discussion of randomised experimentation and meta-analysis in the evaluation of interventions. In Taylor, M. A. P., and Tisato, P. M. (Editors), *Papers of the 27th Australasian Transport Research Forum*. Adelaide: Transport Systems Centre, University of South Australia.

Morrison, D. S., Petticrew, M., and Thomson, H. (2003). What are the most effective ways of improving population health through transport interventions? Evidence from systematic reviews. *Journal of Epidemiology and Community Health*, 57, 327-333.

Mosteller, F., and Boruch, R. (Editors) (2002). *Evidence Matters. Randomized Trials in Education Research*. Washington, D.C.: Brookings Institution Press.

Oakley, A., Strange, V., Toroyan, T., Wiggins, M., Roberts, I., and Stephenson, J. (2003). Using random allocation to evaluate social interventions: Three recent U.K. examples. *Annals of the American Academy of Political and Social Science*, 589, 170-189.

Peterson, A. V., Mann, S. L., Kealey, K. A., and Marek, P. M. (2000). Experimental design and methods for school-based randomized trials: Experience from the Hutchinson Smoking Prevention Project (HSPP). *Controlled Clinical Trials*, 21, 144-165.

Petrosino, A., Turpin-Petrosino, C., and Buehler, J. (2003). "Scared straight" and other juvenile awareness programs for preventing juvenile delinquency. Campbell Review Update I. In: *The Campbell Collaboration Reviews of Intervention and Policy Evaluations (C2-RIPE)*. Philadelphia: Campbell Collaboration.

Rausch, A., Wong, J., and Kirkpatrick, M. (1982). A field test of two single center, high mounted brake light systems. *Accident Analysis and Prevention*, 14, 287-291.

Retting, R. A., Chapline, J. F., and Williams, A. F. (2002). Changes in crash risk following re-timing of traffic signal change intervals. *Accident Analysis and Prevention*, 34, 215-220.

Roberts, I., Kwan, I., and the Cochrane Injuries Group Driver Education Reviewers. School based driver education for the prevention of traffic crashes (Cochrane Review). In: *The Cochrane Library*, issue 1, 2004. Chichester: Wiley.

Sherman, L. W. (Editor) (2003). Misleading evidence and evidence-led policy: Making social science more experimental. *Annals of the American Academy of Political and Social Science*, vol. 589.

Tamburri, T. N., Hammer, C. J., Glennon, J. C., and Lew, A. (1968). Evaluation of minor improvements. *Highway Research Record*, No. 257, 34-79.

Watson, L., Small, R., Brown, S., Dawson, W., and Lumley, J. (2004). Mounting a community-randomized trial: Sample size, matching, selection, and randomization issues in PRISM. *Controlled Clinical Trials*, 25, 235-250.