APPENDIX F

Enhancing Future CMF Research

This appendix was contributed by Dr. Ezra Hauer upon request by the Project 17-63 team.

Introduction

Crash Modification Factors (CMFs) allow one to anticipate the safety effect of manipulations (interventions, design changes etc.). The main aim of this project is to suggest ways to make CMF estimates more trustworthy and more widely applicable.

The essence of CMFs is that they are to capture the safety effect of causes. The commonly used 'before-and-after' and 'cross-section regression modeling' approaches to CMF estimation often run into problems and consensus about trustworthy CMF estimates is slow to emerge. The nature of these problems and some promising directions to alleviate them are examined in the companion paper (Hauer, 2017). In view of the directions identified, what alternative approaches to the determination of the safety effect of causes hold promise? This is the question to be examined.

1. Why Look for Alternative Ways to Get CMFs

CMFs are essential for evidence-based practice. Unfortunately, CMF estimates are at times not well trusted. Some reasons for the mistrust are justifiable, some are not. One may think that CMF estimates obtained in the past, elsewhere, and under different circumstances may not apply here and now. This source of mistrust can be justifiable. It can be remedied, partly, by making CMF's a function of circumstances and time. Occasionally mistrust is of the ‘not-invented-here' variety. It may be difficult to let go of practices and decisions made in the past and which, in the light of more recent research findings, appear to have been inferior or incorrect. This kind of mistrust is not justified. Another source of mistrust may be that CMF estimates are at times all over the place, that the research methods are suspect, and that data were insufficient and of poor quality. When CMF estimates do not agree with each other and it is not clear why, they justifiably do not inspire confidence.

To illustrate the quandary, the ‘revised AMF’ and the ‘Harwood et al’. curves in Figure F1 imply very different crash modification functions[1]. Would reducing the radius of a horizontal curve from 4000’ to 2000’ increase accidents by 10% as per Harwood et al. or by 80% as per Bonneson and Pratt? What the right answer is impacts on practice.

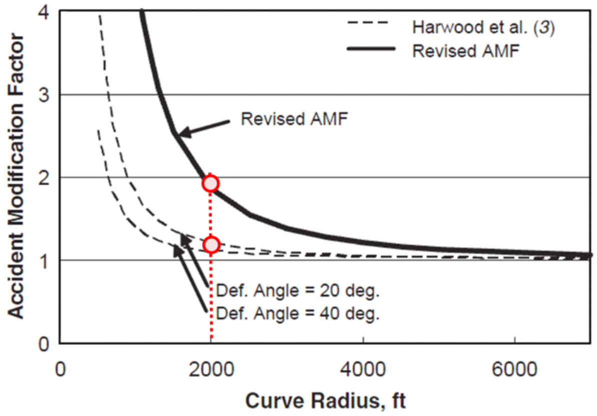

Figure F1. Adapted from Figure 2 in Bonneson and Pratt (2008)

If some of the justifiable misgivings could be alleviated and if CMF estimates could be endowed with more authority and prestige, then practitioners would more likely abide their guidance, and society could

[1] The A in AMF stands for Accident while the C in CMF stands for Crash and so the two acronyms are identical concepts.

be spared some of the unnecessary misery that comes from road accidents. This is why those tasked with road safety management must ask: “what can be done to get trustworthy CMF estimates?”

The proven way to find out what is the effect of a manipulation is by experiments. According to Shadish et al. (2002): *“Experiment: A study in which an intervention is deliberately introduced to observe its effects.”* But in research about CMFs we do not tend to do experiments. When building, maintaining and operating the infrastructure, actions come first and attempts to determine their safety effect are usually an afterthought. Studies of this kind are called ‘observational’[2].

As shown in Hauer (2017), extricating the safety effect of manipulations from observational studies can be difficult. Part I of that report consists of a detailed review of eight studies about the safety effect of pavement marking retroreflectivity (PMR). All eight used observational data. Five studies fitted model equations to cross-section data, one study used five variants of explicitly causal frameworks (also with cross-section data), one study was of the before-after kind and tried to detect the safety effect of PMR in four ways, and one study used monthly time series data about crashes and of PMRs to determine how safety changes as retroreflectivity diminishes. Taken together these studies did not lead to consensus and the question asked was why.

To answer, several problems of method encountered in the course of these reviews were identified. These serve to explain, in part, why progress towards consensus proved to be elusive. The more fundamental reasons for the lack of convergence to a consensus are discussed in Part II. These stem from the limitations of the two most commonly used study prototypes: the observational before-after study (with either comparison or reference populations of units) and the observational study in which a single equation is fitted to cross-section data. The shortcomings of both study prototypes stem from the fact that they are not designed; the researcher scavenges for what data is available after the completion of the projects; the needs of the research are not considered at the time the future project is planned and the manipulation are on the drawing board.

The aim of this report is to make suggestions about some strategic options for future research about the safety effect of manipulations. For future research about the safety effect of manipulations to more surely lead to consensus the following four strategic directions that could be pursued:

1. *In fields not dissimilar to road safety the conduct of randomized experiments is the gold standard and the engine for progress towards evidence-based consensus. In research about the safety effect of manipulations this option is nowadays not entertained. I think that randomization is feasible in some circumstances and that it should be given a fair hearing.*
2. *To attribute cause to effect in one has to account for the effect of uncontrolled nuisance influences. This is easiest to do so accurately when the number of nuisance influences is limited, when they are well measured, and when the function linking them to target crashes is known. These principles should guide us in seeking opportunities for fruitful and trustworthy research; the same principles should serve for identifying those research approaches that are less likely to produce trustworthy results.*
3. *Other disciplines benefited from research approached seldom used in road safety. Specifically, the possibilities embedded in Structural Equations Modeling, in Potential Outcome Causal Inference and in Causal Diagram modeling should be evaluated.*
4. *Central to research about the safety effect of manipulations is the ability to predict what would have happened without the manipulation. The various study types surveyed in Part II of the companion report differ mainly in the way these predictions are produced. Empirical research about which approach to prediction is best is limited. Such a research program is feasible and should be pursued.*

These are the four directions discussed below.

[2] *“An observational study is an empiric investigation of effects caused by treatments when randomized experimentation is unethical or infeasible."* "Paul R. Rosenbaum, Design of Observational Studies. Springer. New York, Dordrecht Heidelberg , London, 2010. Preface.

2. Can We Randomize?

This is what Wikipedia says: *“In science, randomized experiments are the experiments that allow the greatest reliability and validity of statistical estimates of treatment effects."* Most will agree that if in road safety research experimentation and randomization were feasible, the resulting CMFs should be deemed trustworthy and, as a result, practitioners could make more cost-effective choices. But most are also inclined to think, perhaps out of habit and without reflection, that in our line of work there is no opportunity to randomize, no control over which unit gets what treatment. The question is whether staying within this mental box is really necessary? I will argue that there are questions that could (and should) be settled by randomized experiments.

2.1. The experience of others

The problems of trustworthiness of conclusions when based on observational data are of course not confined to road safety. To mention some high profile cases in epidemiology, Rosenbaum (1996) relates the story of the 1976 study by Cameron and (Nobel prize winner) Pauling which, using observational data, claimed vitamin C to be effective for advanced cancer treatment. That claim was refuted in 1985 by a randomized controlled experiment. Similar debacles occurred when observational data indicated that hormone replacement therapy offers some protection from myocardial infarction or that β-carotene protects from lung cancer. Vanderbrucke (2004) refers to these episodes as *“total failures in which results from observational studies were completely overturned by randomised studies.*” As in transportation engineering bad epidemiologic advice has costs; it leads to bad heath policy and medicine.

Another domain in which randomization is rare and observational studies used to be the rule is econometrics; a field about which Imbens (2014) says one tries *“to draw causal inferences in settings where the treatment of interest cannot be credibly viewed as randomly assigned”* and where even after conditioning on additional covariates *“the assumption of no unmeasured confounders does not hold.”* (These quotes seem to also fit the reality of CMF research.) About econometrics Turney (2013) says that in terms of predicting consequences of interventions *“by common consent, they do it amazingly badly.”*

Just as the limitations of observational studies are well known, so is the attraction random assignment to treatment and control. This is why, at this point, attention will shift to the opposite side of the coin – the conduct of experiments and randomization.

Salsburg (2001) relates the story of R.A Fisher’s coming to the Rothamsted Agricultural Research Station early in the twentieth century. The results of ninety years of experimentation with ‘artificial manures’ (i.e. fertilizers) before Fisher’s arrival *“was a mess of confusion and vast troves of unpublished and useless data ... The most that could be said of these artificial manures was that some of them worked sometimes, perhaps, or maybe. ”*(p.5-6) The Rothamsted scientists used various adjustments for differences in rainfall and soil fertility but Fisher showed that *“the year-to-year differences in weather and in ‘artificial manures’ were “confounded” ... there was no way to pull them apart”.* (Also reminiscent of the state of affairs in CMF research.)

For an experiment to lead to useful conclusions one has to be able to say that the effect that is measured is due to the treatment administered and not due to the nuisance influences over the change of which one has no control. This fact lead Fisher to notion of randomization in the allocation of units to either ‘treatment’ or ‘control’. The rest is history.

Disciplines in which conduct randomized trials is accepted and engrained (e.g. medicine) moved consistently and rapidly into their evidence-based historical phase. In other fields in which the allocation of treatment to units is not dictated by research design (sociology, economics) progress was tortuous and the attainment of clarity evasive. Perhaps due to a long-standing perception that such studies gave econometrics a bad name (Leamer, 1983), there is now much interest in natural experiments (in labor economics) and randomized experiments (in development economics) (Imbens, 2010; Angrist and Pischke, 2010).

2.2. Lessons from our own history

As already noted, it is not in our tradition to do experiments the purpose of which is to ascertain what is the safety effect of manipulations. One well known exemption was the "DeKalb project" (Stock et al., 1983). The aim was to evaluate the safety effect of driver training in high schools. The popularity of such training declined sharply after 16,000 high-school student volunteers were randomly assigned to three treatment levels and no significant safety benefit could be shown to exist.

Perhaps of more interest in the present context are two old infrastructure-related studies that can be called experiments; their purpose was to determine the safety effect of edgelining[3].

In 1957 Ohio initiated a program of painting edgelines on rural two-lane roads. Musick (1962) reports the results of an experiment about the safety effect of edgelining. In this experiment nine pairs of road segments were selected so that the roads in each pair were similar. One segment of each pair was selected at random and got edgelines, the other segment served as control. The 'before' period was 1956, the 'after' period was one year after edgelining.

A similar study for Kansas is reported by Basile (1962). In Kansas edgelining also started in 1957 and *"No research had been attempted in Kansas prior to this time to determine the effectiveness of this device or of its economic justification."* (p. 80). After the edgelining program was started some before-after comparisons showed important safety gains. Perhaps to confirm these, perhaps because the edgelined sections *"…had comparatively high initial accident experience rates…"* [4] the randomized experiment based on 1960 data was conducted. Here 29 pairs of adjacent road segments were formed and the decision which segment of the pair gets edgelining was also done at random. I tried to find out what were the historical circumstances surrounding the conduct of these experiments but failed. Even so, some aspects of this long-forgotten episode deserve comment.

In both cases the decision to embark on a program of edgelining preceded the availability of conclusions from the experiments. As Basile attests, in 1957 several states made steady use of edgelines *"based on an anticipated reduction in accidents and fatalities"* (p. 80). The decision to edgeline was based on an 'anticipation' and was unsupported by then available data-based evidence[5],[6].

The experiments were carried out after-the (implementation)-fact and by the same agency that was running the edgelining program. This situated the research in an inhospitable setting. What if the research showed that what was 'anticipated' did not pan out? The results of these experiments serve to illustrate the point.

Both studies found that, following edgelining, the number of accidents at intersections and driveways was much less than what would have been expected without edgelining. This, in itself, was deemed surprising because the mechanism to bring about such an effect was unclear[7]. More surprisingly and contrary to what was 'anticipated', both studies found that away from access points the number of accidents with edgelining was larger than what would have been expected without edgelining[8].

[3] It is possible that more randomized experiments were done only I did not find them.

[4] Possible regression to mean bias.

[5] If the aim of these experiments was not to provide evidence for deciding whether edgelining is worthwhile, what then was their purpose? At least in Kansas, so it seems, the purpose of the experiment was to confirm in a scientifically solid way what preliminary before-after studies indicated; namely, that edgelines reduce accidents.

[6] The habit of implementation based on 'anticipation' and on 'judgment' and without waiting for evidence is deeply engrained in engineering practice. (This is how highway geometric design standards were and are formulated; but this is a different story.)

[7] Basile speculated that with edgelining drivers might be looking further ahead or that the termination of the edgeline before the access point might make its presence more obvious. He suggests that "Carefully planned research is needed to test … these theories." (page 83)

[8] In Ohio *"...the net increase in accidents was approximately 15%..."* (Musick, page 4); in Kansas *"... a 27% net increase in accidents in ... is found"* . However, *"This net change has a low level of significance…"* (Basile, page 83)

Had this been known before the DOTs was committed to and embarked on the edgelining program, it would have been prudent to only paint edgelines in the vicinity of access points and to engage in more research and experimentation. However, inasmuch as the edgelining programs were already running, doing so may have been administratively unpalatable. And so, the authors couched their conclusions carefully so as not to be seen to negate what the employer was doing.[9]

It is difficult to know whether and how these results influenced subsequent practice in Ohio, Kansas and elsewhere. It is clear, however, that the chance of these experiments to be influential was hampered by the two facts. One, that it was too late; the question of whether to place edgelines on roads was already decided and was being implemented. Second, that the researchers were employed by same operating agency that adopted the edgelining program and may have thought it disloyal to emphasize the conclusion that edgelining seems to degrade safety away from intersections and driveways and that the program may need to be modified.

The lesson I draw from this historical episode is that for evaluative research to be useful and influential it must be integrated with the decision-making structure. If the purpose of these experiments was to support the decision on whether and how to edgeline, its results should feed the decision-making process; the experiment should have been completed before the edgelining program begun. If the purpose of these experiments was for others to learn from the experience of Iowa and Kansas, they should have been carried out by an independent agency, perhaps as a part of a national program of CMF research.

2.3. Opportunities for randomization

The two aforementioned experiments were set in what today is an unusual circumstance. It was early days for edgelining and one could still find matching road sections with and without edgelines. Nowadays it would be difficult to find such road sections. Does it mean we cannot do randomized experiments because the era in which it was possible has passed?

I said earlier that randomized experiments should be conducted because they promise to lead to consensus about the safety effect of manipulations. Here I will argue that opportunities to conduct randomized experiments about the safety effect of manipulations still do exist.

Consider e.g., the vexed[10] question whether the repainting faded pavement markings reduces the expected number of target accidents. On the face of it, the conduct of a randomized experiment would be straightforward: Make a list of road segments on which pavement markings should be repainted in the coming season with twice as many segments as is usual. Select at random half of the segment for repainting ('treatment') leaving the other half as 'control'. Interpret the results in the usual manner. If necessary, continue do so for a few pavement-marking seasons till the results are clear.

The safety effect of repainting pavement markings is not the only question that could be settled by a randomized experiments. Operating agencies maintain several lists of projects arranged in order of priority that will be carried out when budgeted. There is list of roads to be resurfaced, a list of intersection to be signalized, etc. Just as with the repainting of pavement markings, every such list is an opportunity for the conduct of randomized experiments.

In addition to prioritized lists and plans for future projects, there are other circumstances amenable to the conduct of randomized experiments. Consider, e.g., the question of the relationship between intersection safety and signal cycle time. One may speculate that the more frequently the signal aspect is changed the larger is the number of accidents. However, road safety reality seldom matches simplistic 'anticipation' and

[9] For Ohio (Musick, page 7) the first conclusion is that edgelining caused: *"...a significant reduction in fatality and injury causing accidents."* which is true because the gain at access points I s larger than the loss elsewhere. For Kansas (Basile, page 86) the first conclusion is that *"...the use of pavement edge markings resulted in a reduction in the number of fatalities."*.

[10] The question is vexed because, as was shown in Part I of Hauer (2017), in spite of repeated research effort it remains unanswered.

one has to test opinion and 'engineering judgment' against data. There are many hundreds of thousands of traffic signals at which, as far as I know, safety considerations do not go into the choice of signal cycle time. Here too the need to know is pressing and experimentation is feasible. If during a certain part of one day the signal network is now using a 60 second cycle time it could be coordinated the next day using a 70 second cycle time. Alternating the two cycle times between days in some appropriately matched and randomized fashion would tell us which cycle time is safer. A related question is the choice of the 'offset' between signalized intersections[11]. One could change the current offsets on randomly selected days. In short, there is a variety of road safety questions that could be answered by the conduct of designed randomized experiments.

There are impediments to such experimentation. My conversation with colleagues from operating agencies points to a reluctance to undertake randomized-controlled experiments. The main sources of this reluctance are three. First, that the conduct of such research is not usually considered to be within the mandate of operating agencies. Second, that experimentation has costs. Thus, e.g., randomization in the repainting of pavement markings would mess up the current routine of contracts and contractors who have been assigned to certain regions and roads. Similarly, if the optimal signal cycle is thought to be 60 seconds then the change to 70 seconds might increase delay. This too is a real cost. Third, there are the ethical and the liability consideration. Is it ethically and legally defensible to postpone the repainting of some pavement markings for one season? Is it ethically and legally appropriate to increase road user delay to learn about the effect cycle time or offset changes on safety? All three objections deserve airing. I will comment on the 'mandate' and 'cost' issues later and discuss here only about the question of ethics.

One could argue, e.g., that not renewing the pavement markings of a road segment that was scheduled to be repainted just so that researchers can have a 'control group' is unethical, and that it exposes the operating agency to the risk of legal liability. This weighty issue deserves study by experts and I am not an expert. We can perhaps learn from medicine, where the conduct of randomized experiments is common and similar issues have already received thought and examination (Friedman et al., 2010).

Two concepts seem salient: 'Equipoise' and 'Ethics Committee'[12]. Initially 'equipoise' meant that doctors can assign patients to one treatment or another only if they believe that each treatment is equally likely to prove superior. But such a notion of equipoise proved to be elusive in practice and in 1987 Friedman (1987) introduced the concept of 'clinical equipoise': a randomized trial is ethical if within the expert clinical community there is no consensus about the relative merits of the alternative treatments. Part I of the companion report may serve as an indication of such a lack of consensus about the safety effect of pavement marking retroreflectivity.

Miller and Joffe (2011) argue that the clinical equipoise criterion is still too narrowly construed because it reflects only the interest of the patients in the trial and *"...ignores the wider societal interest in evidence-based health policy, as reflected in regulatory decisions to approve new treatments for licensing and in health coverage decisions ...."* (Page 476). This debate too needs to be transposed into the road safety context.

The second salient concept that could be adopted and adapted is that of an 'ethics committee'[13], a body that has been formally designated to approve, monitor, and review research involving humans. The purpose

[11] Offset is the time between the onset of green at adjacent signals.

[12] "…is a committee that has been formally designated to approve, monitor, and review biomedical and behavioral research involving humans. They often conduct some form of risk-benefit analysis in an attempt to determine whether or not research should be done.

The purpose of the review process is to assure, both in advance and by periodic review, that appropriate steps are taken to protect the rights and welfare of humans participating as subjects in a research study. A key goal of IRBs is to protect human subjects from physical or psychological harm, which they attempt to do by reviewing research protocols and related materials."

[13] Variably called an institutional review board (IRB), an independent ethics committee (IEC), ethical review board (ERB), or research ethics board (REB).

of such committees is to assure, both in advance and by periodic review, that appropriate steps are taken to protect the rights and welfare of humans affected by a research study[14].

I do not think that the aforementioned impediments (mandate & cost, ethics & liability) fully explain why during the last half century there were no experiments about the safety effect of manipulations. Be it as it may, in the last decade or so, there was a noticeable shift toward evidence-based road safety management. A change of attitude toward the conduct of randomized controlled experiments may now be possible.

2.4. What to do

Randomized experiments are a respected tool for determining the causal effect of manipulations. Even more importantly, they are the engine of progress towards evidence-based practice. I have shown above that the conduct of randomized experiments in road safety is feasible in some circumstances and should be considered.

At present, the question of "what is the safety effect of…." is usually answered by either scavenging for 'all-but-the-kitchen-sink-data' and fitting a model equation to it, or by engaging in an observational before-after study. This practice has not made for quick and steady progress. To do better a change of attitude is needed. It would be better if the usual response to the "what is the safety effect of…." question was "how can I design a study to best answer the question." One option for designing such a study is to conduct a randomized experiment and therefore the task is to determine how to get organized so that randomized experiments assume their proper role in road safety management.

I noted earlier that while only an operating agency can organize randomized experiments it is not its mandate to engage in research the results of which are of benefit to many other operating agencies. The costs of such experimentation are real and a mechanism is needed to share the costs by all those who will benefit. It follows that the initiative for undertaking experiments of this kind and the underwriting of their real costs must come from some kind of umbrella organization that is empowered by many operating agencies (FHWA, AASHTO, NCHRP etc.). That umbrella organization should be responsible for initiating randomized experiments, for paying their cost, and for quality control. Good research results are useful only if they influence practice. Therefore the same cooperative organizational framework should also see to it that trustworthy research results find their way into decision-making, standards, warrants, education and practice.

3. Designing Observational Studies[15]

There are, of course, circumstances in which randomization is not practical or ethical and where information about the safety effect of manipulations must be extracted from observational data. While doing so may be difficult, it is not impossible; one has to be clever about it. In observational studies there are many nuisance influences that can masquerade as cause and cleverness is in the ability to neutralize the impostors. The neutralization of impostors is mainly through study design.

As Rosenbaum (2010) observes, "*The quality and strength of evidence provided by an observational study is determined largely by its design. Excellent methods of analysis will not salvage a poorly designed study* ." Similar points are made by Angrist and Pischke (2010) who say in the abstract that, in

[14] May I add the subversive notion that another ethics committee is needed to protect the rights and welfare of humans affected by highway design and operational decisions.

[15] A book-length coverage is in Paul R. Rosenbaum, Design of Observational Studies. Springer. New York, Dordrecht Heidelberg , London, 2010 and at a less general but more technical level in Rosenbaum, P.R.: Observational Studies (2nd ed.). New York: Springer (2002). Rosenbaum's main concern is with the choice of control groups. The control group serves to predict what would have been the outcome had there been no manipulation. But this is only one of several ways to so predict. In this sense my concern with the design of observational studies is broader.

econometrics" ... *the* <u>*credibility revolution*</u> *in empirical work can be traced to the rise of a design-based approach that emphasizes the identification of causal effects. Design-based studies typically feature either real or natural experiments and are distinguished by their prima facie credibility and by the attention investigators devote to making the case for a causal interpretation of the findings their designs generate."* The same is said by Rubin (2008) in the paper title that *"For Objective Causal Inference Design Trumps Analysis"*. In research about the safety effect of manipulations we too need a 'credibility revolution'.

Progress towards consensus will be steadier and faster if instead of scavenging for after-the-fact-data, research studies could be initiated at the time when the projects that will furnish the data are being planned and designed. What then can be done to design observational studies?

3.1. Desiderata

To attribute the change in the Property of Interest[16] (PoI) to a manipulation, one has to account for the effect of nuisance influences, those which do affect the PoI (Property of Interest) but the effect of which is not the subject of investigation. In Part II of the companion paper I discussed five study prototypes which were arranged in the order of diminishing control over the source of data. The first prototype was a laboratory experiment in which control over the nuisance influences can be near complete. When so, when it is possible to keep all nuisance influences constant, then the measured change in the PoI is clearly due to the manipulation alone.[17] In this case there is no need to account for the effect on the PoI of the change nuisance influences. The second prototype was also a laboratory experiment except that the control over the source of data was diminished; because some nuisance influences did change, a correction had to be applied for the effect of this change on the PoI to be accounted for. This is best done when the nuisance influences are few, when they change little between the 'without' and the 'with' manipulation conditions, when in both circumstances they are accurately measured, and when the function linking them to the PoI is well known. These four desiderata give meaning to the notion of 'study design'. It is desirable to:

(1) reduce the number of nuisance influences;
(2) keep the change or difference in nuisance influences small;
(3) measure their magnitude in the 'without' and 'with' condition accurately;
(4) make sure that the function linking the nuisance influences to the PoI is well known.

3.2. Exploiting cross-section data I: Opportunities and quasi-experiments[18].

Evans (1986) estimated the safety effect of seatbelt wearing by comparing the risk to die of belted and unbelted occupants. Here, as always[19], one can say that a manipulation is the cause of the change in safety when one cannot think of nuisance influences that could have been at work. Thus, one could not just examine the proportions belted drivers or passengers who die in crashes for such a simple comparison of such ratios would be hopelessly confounded by the many differences in crashes, vehicles, age and gender

[16] The Property of Interest here is the safety of units. The safety of a unit is usually defined as the expected crash frequency by severity.

[17] This is in accord with the Second Canon of J.S. Mill (1882)

[18] Shadish et al (2002) say that a quasi-experiment is "*an experiment in which units are not assigned to conditions randomly"*. This is too broad. A better description is *"There are many natural social settings in which the research person can introduce something like experimental design into his scheduling of data collection procedures (e.g., the when and to whom of measurement) even though he lacks the full control over the scheduling of experimental stimuli (the when and to whom of exposure and the ability to randomize exposures) which makes a true experiment possible. Collectively, such situations can be regarded as quasi-experimental designs."* (D.T. Campbell and J.C. Stanley, Experimental And Quasi-Experiment Designs For Research, Houghton Mifflin Company, Boston 1963, p. 34)

[19] That is in accord with Mill's Second Canon.

of those who wore a seat-belt and those who did not. To isolate the effect of seat-belt wearing Evans had to devise a study design that would keep all such nuisance influences constant. To eliminate the nuisance influences of speed, car mass, etc. the comparisons were based on driver and passenger data from the same car and crash. Other potential nuisance influences, those related to gender, age, could be eliminated by using homogeneous age and gender cohorts. The influences of position in car, etc. were accounted for by the Double Pair Comparison Method (Evans, 1991). And so, by being clever about it, in the end the difference in the risk to die could be cleanly attributed to whether seat-belt was or was not worn.

While the seat-belt study is a particularly compelling example of clever design, other opportunities to examine the safety effect of manipulation while holding nuisance influences (nearly) constant have been exploited. Thus, e.g., some studied the safety effect of policing by examining accident frequencies surrounding police strikes. This is a quasi-experiment in which 'pre', 'during', and 'post' strike crash frequencies could be compared on the assumption that the change in the nuisance influences throughout that period was insignificant. Others examined the safety effect of light conditions (or perhaps of sleep duration) before and after the periodic changes in daylight saving time.

Our concern here is with the safety effect of manipulations (design decisions) that, practically, can only be extracted from cross-section data. Examples are road features that are only rarely changed (e.g. the radius of horizontal or vertical curvature, grade, lane width, compass direction, etc.) and therefore cannot be studied using before-after data. The question is how, in this circumstance, can the confounding effect of nuisance influences be minimized; how to be clever about the study design so that they lead to defensible CMF estimates.

One promising approach has been described by Bonneson and Pratt (2008). The idea is to compare road segments along the same road sharing the same traffic, environment, road users and other nuisance influences but differ in the Property of Interest. As Bonneson and Pratt say "*the procedure is based on the use of matched pairs of road segments. The segment pairs … are selected such that their attributes are identical, except for differences in the attributes* (of interest) … *By selecting pairs of matched segments, the effect of the selected attributes on safety is isolated and all other factors are controlled."* (Page 41). They illustrate the procedure by comparing segments of the tangent section preceding a horizontal curve with segments of the same length but located on the curve. This ensures "*similarity in environment as well as geometry and traffic stream*". Thus, e.g., when the crash frequency of a segment on the curve is compared to that on the adjacent tangent, differences in many nuisance influences (driver demography, proximity to hospitals, precipitation, car fleet, blood alcohol content etc.) cannot play a role in the outcome. While there may be other nuisance influences (e.g., differences in the presence of guard rail) that need to be accounted for, the beauty of the idea is in that it eliminates many nuisance influences, even those that are difficult to measure or are unrecognized. Inasmuch as differences in the traffic, environment, and similar nuisance influences are eliminated because of the propinquity of the compared road sections, I will call it the 'propinquity study design'.

Bonneson and Pratt implemented the idea for horizontal curves[20] and show that the same could be done for lane and shoulder width. It seems to me that the idea could also be used for examining how safety varies on crest and sage vertical curves, along sections of grade, perhaps as a function of roadside hazard rating, sideslopes, ditch design, etc. In addition, with the abundance of data in the HSIS and State data banks, one can think of implementing the method state by state, by AADT groups, speed limits etc. thereby making the CMFs a function of circumstances. Doing so would enhance both credibility and transferability. With such implementation additional opportunities may become manifest.

The same basic idea could be implemented in a 'big data' manner. To explain, Bonneson and Pratt preselected a specific CMF, that for the radius of horizontal curves, and then proceeded to look for stretches of road consisting of tangents and curves to be the source of their data. One could reverse the order. The first step would be to scour data sets such as the HSIS or a State data bank to find stretches of road segments

[20] The method can benefit from further development and the need for the calibration of a model deserves examination.

without intersections or major traffic generators. These stretches have the same traffic, environment, maintenance, and many other nuisance influences. Next divide the stretch into shorter segments and determine pair wise, in what traits each pair differs. The difference could be of shoulder type, sideslope, PMR, or any other data element. The crash history of a segment pair that differs in a trait is information to be used to estimate the CMF for that trait difference. By accumulating such information over the HSIS or the State data sufficient information may to estimate many CMFs of interest.

In sum, the propinquity study design seems to go a long ways towards the first two desiderata; that of reducing the number of nuisance influences and that of keeping the differences in nuisance influences small. The approach needs to be tested, developed and refined, and if found promising, widely used.

3.3. Exploiting cross-section data II: Alternative approaches to modeling

In road safety research the most common way to extract CMFs from cross-section data is by fitting to it a single equation. The limitations of this approach were discussed in the companion report. In other disciplines (such as economics, epidemiology and sociology) which face a similar reality, alternative approaches are used. As Karwa et al. (2011) note: *"The research questions that motivate transportation safety studies are causal in nature. Safety researchers typically use observational data to answer such questions, but often without appropriate causal inference methodology. The field of causal inference presents several modeling frameworks for probing empirical data to assess causal relations.*" (Abstract, emphasis added). The potential of these presently unused causal modeling frameworks to produce CMFs deserves attention.

One of the obstacles to the causal interpretation of a single equation model fitted to cross-sectional data is the absence of an underlying structure and theory; the model equation does not specify what trait (variable) influences what and how and the place of prior knowledge is taken by an assumed simple function. Statistical techniques are good at estimating the values of unknown parameters but are not good at discovering the structure of the functional relationships between variables[21].

In our road safety data, there is plenty of structure. Consider, e.g., the simple causal diagram in Figure F2. Several questions can now be asked.

[21] On this I say in the 'Art of regression modeling in road safety' (page 93) that " *Even leading advocates of the possibility to interpret observational data causally take care to distance themselves from single-equation regressions. Thus, e.g., Bollen and Pearl* (2013) *carefully differentiate between Structural Equation Models (SEMs) and regressions. The main distinction between the two is that in an SEM the researcher must specify 'what causes what' such that "each equation (in the SEM) is a representation of causal relationships between a set of variables, and the form of each equation conveys the assumptions that the analyst has asserted" (p.4). Furthermore, that these causal assumptions "derive from prior studies, research design, scientific judgment, or other justifying sources" and that "(t)he analysis is done under the speculation of 'what if these causal assumptions were true.'" (p. 9).*

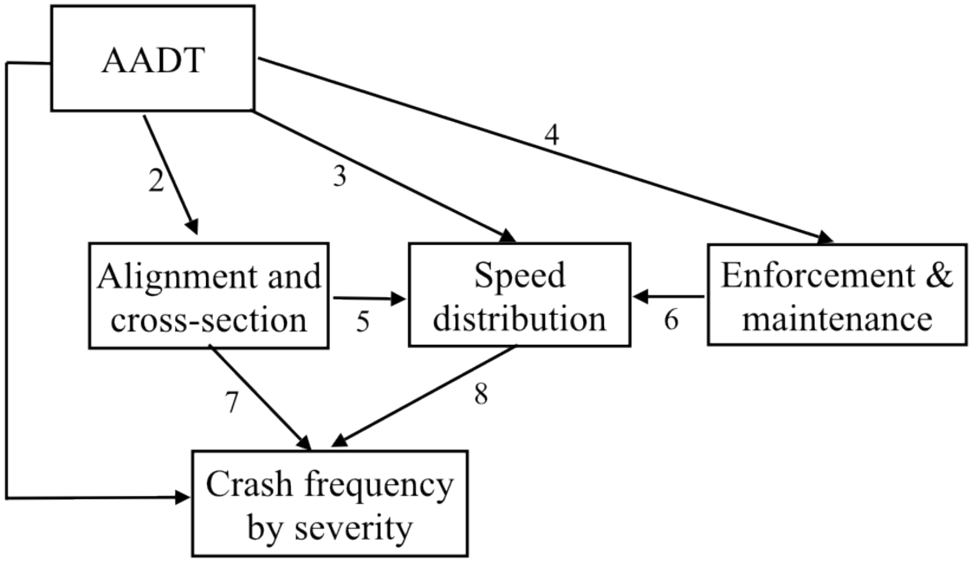
AADT
1
2
3
4
Alignment and cross-section
5
Speed distribution
6
Enforcement & maintenance
7
8
Crash frequency by severity

Figure F2. A simple causal structure

In this figure the safety of a road (the 'Crash frequency by severity' node at the bottom) is influenced by the various alignment and cross-section elements via causal arrow 7 and also by traffic via arrow 1. But the forecasted traffic flow was what influenced the road alignment & cross section during the planning stage via arrow 2. To predict the change in safety due to a change in AADT one might be interested in its effect via arrow 1 and perhaps through arrows 3 and 8 but excluding its effect via arrows 2 and 7. How can these influences be separated?

The causal arrows represent relationships about which we have some prior knowledge. We may have an adequate empirical relationship to attach to arrow 3, perhaps to arrow 8, but lesser knowledge about arrow 1. How should prior knowledge be incorporated in our modeling?

To tease out of such a causal diagram the safety effect pavement marking retroreflectivity one might have to split 'enforcement' from 'maintenance' and link the latter by causal arrows directly to the crash frequency and severity node and also to the 'speed distribution' node. Should and could the direct and indirect influences by separated?

It is obviously unwise to disregard the structure which we know to exist within our data, as it is unwise to do modeling without making use of what is known from prior research about the relationship that accompany the causal arrows. There is long-standing and extensive experience with causal network modeling in other fields but I know too little about causal diagram and structural equation modeling to make specific and solid recommendations. However, it is perhaps self-evident that the experience of others with such causal modeling should be evaluated and its promise for the estimation of trustworthy CMFs evaluated.

4. What is the Best Way to Predict?

The aim of CMF research is to determine what change in safety is caused by manipulations[22]. To determine what change in safety is caused by the manipulation of some trait of interest, one must compare what safety was with manipulation to *what safety would have been at the same time but without the manipulation*. What safety was with the manipulation is based on observation, measurement and statistical estimation. What it would have been at the same time but without the manipulation cannot be observed and measured since such a state never existed; it can only be predicted with varying degrees of plausibility and confidence. The "at the same time" phrase makes it clear that safety-related traits unaffected by the manipulation are the same when the 'what was' and the 'what would have been' are compared. This essential comparison between what can be estimated and what can only be predicted is shown in Figure F3.

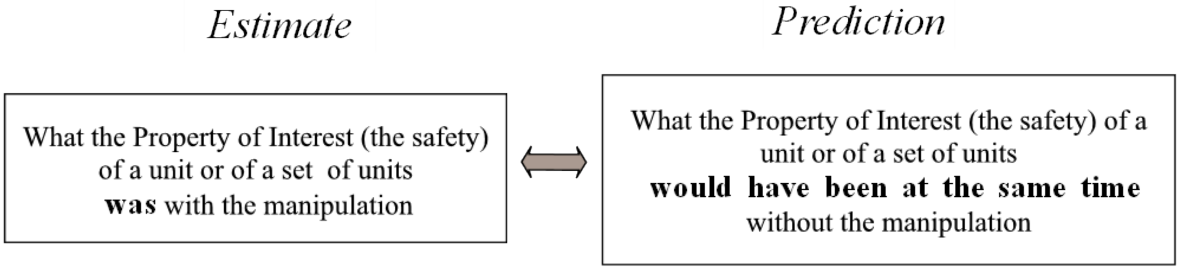
Estimate
Prediction
What the Property of Interest (the safety) of a unit or of a set of units **was** with the manipulation
What the Property of Interest (the safety) of a unit or of a set of units **would have been at the same time** without the manipulation

Figure F3. The essential comparison (Figure 14 from Hauer, 2017).

One has to predict 'what would have been' irrespective of whether the data are of a before-after or a cross-section kind. To illustrate, in a before-after setting, say, if a stop-sign controlled intersections is signalized, then one has to predict what in the after signalization period would have been the safety of this intersection had it remained stop-controlled. Alternatively, in a cross-section setting, where one is comparing stop-controlled and signal- controlled intersections, one has to predict what would have been the safety of the signalized intersections if they were stop controlled.

There are many ways to predict. One can predict by extrapolation of past trends, by use of comparison groups, by representation of known causal relationships, by hybrid methods of prediction (e.g. Empirical Bayes method), etc. In addition, some of these 'ways' come in several variants. Thus, e.g., one can extrapolate past trends by using 1,2,…. n most recent observed crash counts, by fitting one of many functions to these data points, by single or double or triple double exponential smoothing or by similar algorithms, by recombining separate extrapolations of risk and exposure, etc.[23] Similarly, if prediction is based on the use of comparison groups, these can be selected by judgment, by trait-based or by propensity score matching, by similarity in time series of crashes, etc. Entire books were written about the many ways of choosing comparison groups (Rosenbaum, 2010).

Even when applied to the same data, the many different ways to predict produce diverse predictions and this diversity translates into a corresponding diversity in what we estimate to be the safety effect of manipulations, the CMFs. The obvious questions is which of the many ways to predicts is best in what circumstances. The quality of the CMF estimates depends on an answer to this question.

One might assume that the question of "how best to predicts what would have been" was , over the years, a prominent theme of road safety research. This does not seem to be the case. I first examined the issue in 1991 (Hauer, 1991), Quaye (1992) picked up the thread between 1992 and 1994, and I returned to it in

[22] The word 'manipulation' is used here in a generic sense, representing words such as 'treatment', 'intervention', 'design alternative', 'change of trait' etc.

[23] I tried some of these in Hauer (2010) pp. 1111–1122.

Hauer, 2010. However, as far as I know, the issue has never been addressed in a comprehensive manner and is far from ready for drawing conclusions.[24]

Even simple questions have not been well examined. Would one predict next year's crashes better by averaging 1 or 2 or 'n' past crashes? (For some reason we tend to use n=3 years as an ad hoc compromise between statistical reliability and historical relevance).More troubling is the fact that the performance of the EB approach has not been empirically substantiated. (I know of only one paper which attempted to do so and the results were neither conclusive nor very convincing). While the logic of EB is sound, much depends on how well matched are the comparison group and the unit for which we are predicting. Similarly troubling is the lack of testing for the HSM method of prediction; the combination of base models and CMF correction factors. Would one not predict better using the full model?

Success in future research about CMFs (whether F stands for Factor or Function) hinges on resolving the question how to predict best in what circumstances. Answering this question should be straightforward. There are neither methodological impediments nor is there a shortage of data. The method for determining which of two approaches to prediction is better has been described (see, e.g., Hauer, 2010, cited in footnote **Error! Bookmark not defined.**). Data are plentiful and readily available for many thousands of units (road segments, intersections, ramps, crossings, etc.). For each such unit we can compare what the crash count was to what it would be predicted to be based on earlier data. All that is required is a comprehensive program of research.

5. About the Integration of Research and Practice

The task of NCHRP 17-63 is to suggest ways for research to produce trustworthy and transferable Crash Modification Functions. The hope is that when CMFs are trustworthy and transferable they will have an easier time to percolate into practice. And so, the implicit and more encompassing aim of this project is the integration of research and practice.

This aim has at least two aspects. One is the need to facilitate the flow of evidence from research into practice. On this there are clear signs of dysfunction and concern about it should give pause (Hauer, 2015). It would be a stretch to argue that this concern about how CMF research influences practice is a part of NCHRP 17-63. Still, a recommendation that the subject deserves high-level consideration (perhaps as a sequel to the effort that produced Special Report 292[25]) might be timely.

The other aspect of research-practice integration is the need to make CMF research an accepted part of what operating agencies help to do. This need came to the fore twice. Once when I suggested that the conduct randomized controlled studies is both important and feasible, that only operating agencies can carry them out, but that they are naturally disinclined to do so. Why would operating agencies help with the conduct of a randomized controlled study when it is not their mandate? That the very idea seems so impractical is testimony to the separateness of research and practice. The other time the need for integration surfaced was when I suggested that observational studies are likely to produce trustworthy results when they are designed at the time that infrastructure and operations projects are on the drawing board and not an afterthought. This too requires attitudes to change and arrangements that makes help with CMF research a standard part of what operating agencies do.

I have been associated with various aspects of CMF research as an academic researcher, as a member of TRB special studies teams, as a contractor, on NCHRP panels etc. Based on this long time experience my impression is that CMF research materializes in a complex environment that evolved gradually over many

[24] Even the EB approach to predicting now popular in before-after studies is based mostly on its theoretical ability to account for regression to mean bias and to reduce the variance of the prediction when the perfect reference group is available than on an empirical determination of it actually outperforming other methods of prediction. It is possible, nay likely, that when the threat of selection bias is small and the reference group imperfect other methods of prediction perform better.

[25] The recommendations of Special Report 292 (TRB, 2008.) were not particularly influential.

years in response to circumstances as they arose, and without being fashioned by a vision or an articulated plan. The extant strands of CMF research practices, taken together, exist in (or make up) some semi-independent domain of activities the ties of which to practice are unspecified and unplanned. While there are many links between CMF research and the world of practitioners and operating agencies, these links too came into being more through ad-hoc responses and arrangements of convenience and happenstance than by premeditation. There is little point in continuing to maintain and foster the research-practice solitudes. It seems to be high time to give consideration to a purposeful integration of road safety research and practice.

6. Enhancing Future CMF Research: A Summary

In the companion report (Hauer, 2017), I argue that continued attempts to extract reliable CMFs by fitting a single-equation models to cross-section data are unlikely to bring about consensus; that to attribute effect to the cause of interest one strive to create conditions in which nuisance influences are minimized and well accounted for. In this report I examine several concrete moves that may be considered and then make a general observation.

The concrete steps are:

1. Randomized controlled experiments are the engine of progress towards evidence-based practice. Conduct randomized controlled experiments. I identified some circumstances in which doing so is possible; other circumstances may exist.

2. Develop and test the propinquity study design based on the Bonneson and Pratt idea. Should it offer promise, establish a research program to widely implement it so as to exploit available cross-section data.

3. Causal inference methods have seldom been used in road safety research but are common in other disciplines facing a similar reality. Establish a research program the aim of which is to study, examine and adapt the various causal inference methodologies used in epidemiology, sociology and econometrics in order to evaluate their promise for the estimation of trustworthy CMFs.

4. The success of CMF research depends on how well one can predict what would have been the safety of units had they been not manipulated. Many methods for generating predictions are in use but it is not clear which performs best in what circumstances. Initiate a comprehensive program of research to find out.

In conclusion a more general observation is in order. All would agree that research about CMFs has value if its results are trustworthy and used in practice. A closer integration of CMF research and practice would enhance trustworthiness of research results and promote their use in practice. A way must be found to make research a part of the live tissue of operating agencies and thereby to break down the barriers separating the two solitudes.

References

Angrist, Joshua D., and Jörn-Steffen Pischke. 2010. "The Credibility Revolution in Empirical Economics: How Better Research Design Is Taking the Con out of Econometrics." Working Paper 15794 National Bureau of Economic Research. (also Journal of Economic Perspectives, 24(2): 3-30.

Basile, A. J. "Effect of Pavement Edge Markings on Traffic Accidents in Kansas." Highway Research Board Bulletin 308, 1962, pp. 80-86.

Bollen KA and Pearl J, Eight myths about causality and structural equation models. In: Handbook of Causal Analysis for Social Research, S. Morgan (Ed.), Springer, 2013.

Bonneson, James A. and Michael P. Pratt, Procedure for Developing Accident Modification Factors from Cross-Sectional Data. Transportation Research Record: Journal of the Transportation Research Board, No. 2083, Transportation Research Board of the National Academies, Washington, D.C., 2008, pp. 40–48.

Evans, L. Traffic Safety and the Driver, Van Nostrand Reinhold, New York, 1991, pages 20-22.

Evans, L.: The effectiveness of safety belts in preventing fatalities. Accid. Anal Prev. 18, 229–241 (1986)

Freedman, Benjamin, Equipoise and the Ethics of Clinical Research. N Engl J Med 1987; 317:141-145.

Friedman LM, Furberg CF and DeMets, D, Fundamentals of Clinical Trials, Springer 2010, in particular Chapter 2.

Hauer, E., An Exemplum and its road safety morals, 2015. Downloadable from ResearchGate

Hauer, E., Ng, J.N.C., Papaioannou, P., 1991. Prediction in road safety studies: an empirical inquiry. Accident Analysis and Prevention 23 (6), 595–607 and Hauer, E., 1991. Comparison groups in road safety studies: an analysis. Accident Analysis and Prevention 23 (6), 609–622.

Hauer, Ezra, Developing consensus in research about the safety effect of manipulations, NCHRP 17-63 Final Report, Appendix G, 2017.

Hauer, Ezra, On prediction in road safety. Safety Science 48 (2010) 1111–1122.

Hauer, Ezra, Retroreflectivity and Safety: Lessons of Past Research. Draft, 2015-12-26.

Imbens, Guido W., Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009), Journal of Economic Literature 48 (June 2010): 399–423, http:www.aeaweb.org/articles.php?doi = 10.1257/jel.48.2.399.

Imbens, Guido W., Instrumental Variables: An Econometrician's Perspective. Statistical Science, 2014, Vol. 29, No. 3, 323–358.

Karwa, Vishesh, Aleksandra B. Slavkovic and Eric T. Donnell, Causal inference in transportation safety studies: Comparison of potential outcomes and causal diagrams. The Annals of Applied Statistics 2011, Vol. 5, No. 2B, 1428–1455.

Leamer, E.E., Let's Take the Con Out of Econometrics. The American Economic Review, Vol. 73, No. 1. pp. 31-43, 1983

Mill, John Stuart. A System Of Logic, Ratiocinative And Inductive, Being A Connected View Of The Principles Of Evidence, And The Methods Of Scientific Investigation. Eighth Edition. New York: Harper & Brothers, Publishers, Franklin Square. 1882, page 483.

Miller, FG. and S. Joffe, Equipoise and the Dilemma of Randomized Clinical Trials. N Engl J Med 364;5, February 3, 2011.

Musick, J. V. "Effect of Pavement Edge Marking on Two-Lane Rural State Highways in Ohio." Highway Research Board Bulletin 266, 1962, pp. 1-7.)

Quaye, K.E., 1992. Forecasting models in road safety studies. PhD Dissertation, Department of Civil Engineering, University of Toronto

Quaye, K.E., Hauer, E., 1993. The use of forecasting models in the evaluation of safety interventions: a theoretical inquiry. In: Daganzo, C.F. (Ed.), Transportation and Traffic Theory. Elsevier Science Publishers, pp. 313–332

Quaye, K.E., Hauer, E., 1994. Assessing forecasting methods used in before after studies. Paper presented at the 72nd Annual Meeting of the Transportation Research Board, Washington, D.C.

Rosenbaum, P.R., Design of Observational Studies. Springer. New York, Dordrecht Heidelberg, London, 2010.

Rosenbaum, Paul R., Observational studies. Springer Verlag, New York, 1996, in section 1.2.

Rubin, Donald B., 2008. For Objective Causal Inference, Design Trumps Analysis'. The Annals of Applied Statistics 2008, Vol. 2, No. 3. 808-840.

Salsburg, David, The lady tasting tea, Henry Holt and Company, New York, 2001.

Shadish, William R., Thomas D. Cook, Donald Thomas Campbell, Experimental and Quasi-experimental Designs for Generalized Causal Inference. Houghton Mifflin, 2002.

Stock, J. R., J. K. Weaver, H. W. Ray, J. R. Brink, and M. G. Sadof. 1983. Evaluation of Safe Performance Secondary School Driver Education Curriculum Demonstration Project. Washington, DC: U.S. Department of Transportation, National Highway Traffic Safety Administration.

TRB Special Report 292: Safety research on highway infrastructure and operations. Transportation Research Board, 2008.

Turney, Jon A model world. In economics, climate science and public health, computer models help us decide how to act. But can we trust them? December 2013. http://aeon.co/magazine/science/should-we-trust-scientific-models-to-tell-us-what-to-do/. Downoladed on 25/11/2014.

Vandenbroucke, Jan P, When are observational studies as credible as randomised trials? The Lancet; May 22, 2004; 363, 9422, pg. 1728.