Annex B. Before and After studies

B.1 The Before and After studies (‘the studies’) source of evidence presented in the *Stocktake* looked at stretches of motorway where a form of smart motorway has been introduced to provide a ‘like for like’ comparison. It typically involved constructing a counterfactual (an estimate of what would have happened if a stretch of motorway had not been converted) which is key to understanding the safety impact of converting it to all lane running (ALR).

B.2 This annex explains our assessment of Highways England’s overall approach, examines the two most significant factors (the counterfactual and statistical significance) in more detail and draws together our overall findings.

B.3 We discuss the statistical properties of data on collisions, casualties and fatalities. This involves using terms like ‘random’ or ‘not significant’ which have a specific meaning in statistics. In using these terms we are not meaning to diminish the importance or significance, in the more general sense, of every road collision, casualty or fatality.

Background

B.4 The studies were considered the strongest source of evidence in the *Stocktake*, which concluded (in paragraph 5.3) that for “specific roads which have been converted to ALR: the overall casualty rate declines significantly; the fatal and serious casualty rate increases slightly, but within the statistical margin of error; and the FWI rate declines.” This is what we are referring to in any references to the *Stocktake’s* conclusions later in this annex.

B.5 This section of our review focuses on the data and evidence underpinning these conclusions and were relevant to all four questions posed in the remit. In relation to reliability, robustness and appropriateness we considered the extent to which the approaches employed were in accordance with HMT Green Book and Magenta Book guidance on public policy evaluation.

Summary of evidence

B.6 We reviewed each of the key pieces of evidence underpinning the conclusions drawn in the *Stocktake*, namely:

- the before and after analysis in the *Smart Motorway All Lane Running Overarching Safety Report* (the *Overarching Safety Report*), including the detailed spreadsheet containing the calculations that were used;
- Post Opening Project Evaluation (POPE) reports for ALR and Dynamic Hard Shoulder (DHS) schemes; and
- the M25 three-year monitoring reports (J5-7 and J23-27).

B.7 In addition, we undertook a high-level literature review of similar types of evidence available internationally.

B.8 The advice and challenge from our independent econometrician was critical to our work.

Assessment of overall approach adopted

B.9 The key strength of a like for like comparison is that it provides a better indication of the safety impact of the changes made to the specific roads. The strength of the evidence depends on the approach and data used.

B.10 The *Stocktake* considered a range of evidence from different sources:

(a) ALR POPE evaluations and monitoring reports – comprising two one-year POPEs and two three-year monitoring reports. This is a relatively limited pool of evidence, with mixed results from which it is difficult to draw firm conclusions.

(b) *Overarching Safety Report* – additional, safety-focused analysis of all nine ALR projects for which post-opening data were available, two with three years of after data and seven with one year. The analysis was carried out identically on all projects, allowing the data to be pooled. This maximised the use of the data but also had some potential drawbacks in terms of masking scheme-specific variation in averages (although scheme-specific results were still available) and imposing implicit assumptions around the comparability of schemes, and the independence of the numbers of collisions on them.

The analysis found statistically significant reductions in total casualties (section 3.3.1) and collisions (section 3.2.1) for ALR as a whole. Highways England also report a seemingly large reduction in the fatal and weighted injuries (FWI). However, this was not tested for statistical significance (which Highways England explained was because of the weightings applied in its calculation); changes in people Killed and Seriously Injured (KSI), which were statistically tested, are either inconclusive or statistically insignificant; and there is no separate analysis of fatalities reported.

(c) DHS POPEs and monitoring reports – the *Stocktake* covers five POPEs and one three-year monitoring report. As with the ALR reports, the evidence is mixed and this is not aided by methodological differences (around the counterfactual and significance testing) between the POPEs and the monitoring report. The *Stocktake* also included additional analysis of eight DHS projects. This could have acted as a useful comparator to the overarching ALR analysis. But its usefulness in this regard is limited by methodological differences – mainly the lack of a counterfactual and significance testing.

B.11 All of the strands of evidence follow an approach similar to how safety impacts are considered in Highways England's long-established programme of POPE analysis. The *Stocktake* states that this is "in line with best practice and HMT guidance on policy evaluation, as set out in the Green Book and Magenta Book." The Magenta Book describes principles and processes to support good quality policy evaluation. The core principles are that the analysis should be useful, credible, robust and proportionate.

B.12 Of these, robustness is central to the scope of this review and the guidance recognises that there is no objective standard against which to judge robustness. The description of the proportionality principle indicates that factors such as a policy's profile and uncertainty might drive decisions around the "scale" of analysis. One of the key factors described under usefulness is the timeliness of evidence. This introduces a trade-off with robustness that is evident in some of the decisions Highways England had to make in its analysis. For example, pooling scheme-level data in its overarching ALR analysis rather than waiting until more years of scheme-level after data were available. Finally, transparency and independent input are important elements of the credibility principle. While Highways England might not have had sufficient time to seek independent oversight of its before and after analysis that supported the *Stocktake*, its credibility could have been enhanced by publishing a clear explanation of its method, and the assumptions that underpin it. The Magenta Book also describes a wide range of different methods but is not prescriptive about which must be used in specific circumstances.

B.13 In Magenta Book terminology, a POPE study is a combination of an impact evaluation (what difference has an intervention made?) and a value for money evaluation (is the intervention a good use of resources?). It has a broad scope looking at many impacts of a project, of which safety is one part. By comparison, our review is focused solely on safety and not on wider value-for-money considerations. Possibly because it is only one of many impacts considered, the POPE approach to safety is a relatively simple application of a 'difference in difference' approach, where a counterfactual is constructed to compare the 'after' scenario with what might have happened without the project, rather than simply the 'before' period. Highways England used similar methods in its monitoring reports and the *Overarching Safety Report* to maintain consistency with the POPEs.

B.14 The *First Year Progress Report* states that increasing motorway capacity (through conversion to ALR) improves safety by attracting traffic from other, more dangerous roads. During the course of this review, Highways England explained how this assertion is based on its standard business case development methods, using high-level collision or casualty rates by road type. POPE analyses, both those already completed and still to be undertaken, represent an opportunity for Highways England to collate outturn evidence to further substantiate this claim.

B.15 The *Stocktake* considered a wide range of sources, each reporting results for different metrics and often with mixed or inconclusive results. The *Stocktake's* conclusions (as quoted at paragraph B.4 of this annex) indicate that more weight was put on the results of the overarching ALR analysis than other elements of the before and after analysis.

B.16 There are methodological differences between some of the POPE and monitoring reports. For example, the one year after POPE for the M25 J23-27 project used five years of 'before' safety data, looking across the "whole modelled area" (not just the converted sections of motorway) that had been used in the business case. The third year monitoring report for the same project used three years of before data and looked only at the upgraded links. Within the overarching ALR analysis, each ALR scheme was treated consistently, eliminating any methodological differences like those illustrated above. This means that the results could be easily combined, or pooled, to look at the overall impact of ALR. Highways England did this by summing the traffic, casualties and collisions for each project and treating the overall numbers as if they were a single project. This is a straightforward approach that helped the company to overcome the small amounts of data available for the schemes individually. However, as discussed above, this approach does not come without its limitations and Highways England did not set out in its reports the assumptions required, or how appropriate they were, under this approach.

B.17 One such assumption, required for the statistical tests of differences, is that each collision or casualty occurs independently of one another. For casualties, this assumption is unlikely to hold, for example, because of variation in the number of casualties per collision. This leads to ‘overdispersion’, or greater variability in the data than you would normally expect. This does not mean that Highways England’s approach is invalid, but rather, the before and after comparisons of collisions should be considered more robust than that of casualties.

B.18 Similarly, just as casualties on a stretch of ALR motorway might not occur independently of one another, the number or rate of collisions on different stretches of ALR motorway might not be independent of one another. There is a risk that individual schemes all experience a ‘good’ or ‘bad’ year (for example, because of a common factor like the weather) which does not appear statistically significant for individual schemes but, because of the larger sample size, does appear significant when the schemes are pooled.

B.19 We have not tested for correlation between the schemes as part of this review and, to some extent, the effect could be mitigated by how the counterfactual is constructed. So, we view this as a potential risk rather than a demonstrated shortcoming of the approach. We also recognise the robustness-usefulness trade-off that Highways England faced in its overarching analysis and consider that pooling scheme-level data together was the right thing to do at the time (albeit we also consider there should have been a clear description of the underpinning assumptions, to aid credibility). We also consider that updating the analysis with more years of ‘after’ data would improve robustness by providing more scheme-level insight and reducing the risk that an anomaly in a single year’s data would skew the pooled, ALR-level results.

B.20 The conclusions in the *Stocktake* quoted at the beginning of this annex describe changes in casualty rates. It is not clear to us to what extent this is based solely on the casualties analysis (which is less robust than that of collisions), or on the analysis of casualties and collisions taken together, especially as the conclusions from both strands of evidence are broadly similar. The *Stocktake* conclusions also foresaw that the “evidence base has the potential to evolve, as more years of data become available and analytical methods develop.”

B.21 While the *First Year Progress Report* was completed to relatively short timescales, we consider that Highways England could have anticipated the need for updated overarching before and after analysis and included it in that report. **We recommend** that Highways England should update the project level and pooled before and after analysis reported in the *Overarching Safety Report* with outturn data for 2019 and 2020 (subject to data for that year being suitably comparable).

B.22 We consider that the approaches to defining the counterfactual and to statistical testing are the most critical to the robustness of the before and after analysis and the conclusions that can be drawn from it. In the following sections we describe Highways England's approach in these two areas and recommend improvements that could be made when the analysis is updated.

The counterfactual

B.23 In an ideal world any before and after comparison would examine the experience of two otherwise identical motorways; one converted to ALR and one operating conventionally. In practice, a 'counterfactual' scenario needs to be constructed that represents the best view of what would have happened if the intervention had not taken place. The purpose of constructing a counterfactual is to control for other factors that affect safety performance that are not the direct result of the ALR schemes. Therefore, the counterfactual is key to the robustness of conclusions that can be drawn from the analysis.

B.24 Highways England employs a counterfactual as a standard part of the safety evaluations included in its POPE reports for all major schemes. This approach has been replicated in the *Overarching Safety Report*. With respect to ALR, the headline conclusions of both the *Overarching Safety Report* and the *Stocktake* itself, are based on a comparison of the post-implementation outcomes and the counterfactual.

B.25 As discussed above, the additional before and after analysis of DHS schemes included in the *Stocktake* did not include a counterfactual. In our view, using a counterfactual is the right approach to before and after safety evaluation. Assessing DHS and ALR on the same basis in the *Stocktake* would have been a better approach and would have provided a valuable comparison point for the performance of ALRs.

B.26 The remainder of this section deals with specific issues related to the way the counterfactual has been constructed.

Construction of the counterfactual

B.27 Highways England's approach (as described in the *Overarching Safety Report*) is to "assume that, if the scheme had not been built, the number of collisions on the roads in the study area would have changed at the same rate as they did nationally [on motorways] during the same period." Highways England is attempting to account for the exogenous factors, for example improved vehicle safety, that affect accident rates that are not related to the changes to the physical layout of the road.

B.28 In practice, the national rate of collisions will include the effect of newly constructed road schemes, including the ALR schemes themselves. This point is noted in the *Overarching Safety Report*. ALR schemes make up a relatively small proportion of the network and we agree with Highways England's assessment that their influence on the national rate of collisions will be very small.

B.29 In recent decades, the underlying rate of collisions on the highway network in most countries has steadily fallen. This has been attributed to exogenous factors such as changes in vehicle technology or behavioural changes. In this context it is understandable that Highways England has developed a method that is directed towards establishing this underlying trend. Otherwise, there is a risk that a reduction in the 'after' collision rate is wrongly attributed to an intervention when, in practice, it is a product of other factors.

B.30 The *Overarching Safety Report* covers the period 2010 to 2018. During this time, the rate of total collisions continued to decline steadily, which highlights the value of the counterfactual adjustment. However, the rate of fatal and serious collisions (KSIs) fluctuated with no obvious overall trend and has been affected by the introduction by some police forces of the CRASH and COPA recording systems.

B.31 Because the nine ALR schemes were constructed over different time periods, the counterfactual adjustments are specific to each scheme. In respect of KSIs, the counterfactual adjustment is positive in some cases and negative in other cases. The results are highly sensitive to the time-period being considered. To illustrate this, the adjustment factor for KSI casualties from 2011 to 2015 is 99% and from 2012 to 2016 is 121%.

B.32 Highways England's approach to the counterfactual is simple to understand and easy to replicate, allowing it to take a similar approach to all its scheme evaluations. However, it has two main drawbacks. Firstly, there are a variety of local factors, such as the weather, the rate of traffic growth, or changes in traffic mix (e.g. the proportion of heavy goods vehicles), that influence collision rates. As a result of these factors, the national rate may not be reflective of the underlying trend in any given area. Secondly, as illustrated in the example above, because collision rates fluctuate from year-to-year, the adjustment factor can be very sensitive to the precise years chosen, and this is particularly the case for KSIs.

Alternative approaches to the counterfactual

B.33 Drawing on the international literature review (see B.500 onwards), we have identified three alternative ways in which Highways England could have approached the counterfactual.

B.34 Firstly, it could have selected a control group, a specific set of comparator motorway links. In theory, if a comparator can be selected with similar characteristics to the road being tested, the resultant counterfactual adjustment factor is likely to be more accurate. Whilst this approach has its merits, it is both time consuming and vulnerable to criticisms of 'cherry picking'. It is also the case that relying on a single or small sample of comparator links is more prone to seemingly random fluctuations than a broader-based average.

B.35 The second approach is to refine how the counterfactual adjustment is calculated by using a more disaggregated rate that better reflects the underlying trend for road types and regions. Highways England has developed a new approach along these lines and employed it in the POPE evaluation of the M3 J2-4a ALR project. We think that Highways England's new approach represents an improvement to the method used in the *Overarching Safety Report*. We also support the concept of identifying a confidence interval, or sensitivity testing a range, around the counterfactual to reflect the uncertainty involved in its calculation, and/or how trends could be smoothed to avoid year-on-year fluctuations artificially impacting on the adjustment applied. **We recommend** these developments to the counterfactual are incorporated into the updates to the overarching analysis as well as in future POPE analysis.

B.36 A third approach that we have observed from the international literature review is to develop simple regression models (often termed safety performance functions – see section B.73) for major road types that seek to establish an expected rate of collisions controlling for changes in factors such as traffic levels and composition. Models can be updated regularly to account for changes in accident rates over time or include a time trend to capture the influence of exogenous factors such trends in vehicle technology. Modelling the counterfactual in this way offers a potentially more sophisticated approach, although we recognise the risks that would have been involved in attempting to develop such tools for the purposes of the *Stocktake*. We also recognise that such approaches are data intensive and would take time to develop. Nevertheless, in the medium-term we think these methods have the potential to enhance the quality of Highways England's safety analysis. **We recommend** that Highways England reviews techniques employed elsewhere, including safety performance functions, and consider whether they provide feasible options for improving safety analysis of the SRN.

B.37 In summary, Highways England's use of a counterfactual adjustment was consistent with the standard approach that has been established in the organisation. It was appropriate that the conclusions of the *Stocktake* were focused on comparing against the counterfactual, rather than a simple before and after comparison which fails to account for underlying trends. Highways England applied its default approach, prioritising consistency with previous analysis, rather than considering the range of alternative methods. Whilst the default approach was broadly sensible, in our view, there are more sophisticated methods available that could improve the accuracy and robustness of the before and after analysis.

Statistical testing

B.38 The number of road traffic collisions varies from day-to-day and year-to-year. Some of this variation might be down to specific factors, such as the weather, but much of it is down to natural variation, or randomness, that occurs commonly in real world data. Statisticians might consider whether this variation fits a particular type of pattern or distribution.

B.39 A DfT technical note published alongside *Reported road casualties in Great Britain* showed how the number of fatal road collisions that occur each day could be characterised by a statistical distribution known as a Poisson. This distribution expresses the probability of a number of events occurring in a given time period, given a known average rate of occurrence and that the events are independent of one another. So in this context, the Poisson distribution could tell you how likely it would be for two fatal crashes to occur in a year on a stretch of motorway if the average is one per year.

B.40 In this instance, statistical tests were used to establish that observed road safety data follow a Poisson distribution. With this established, statistical tests can also be used to attempt to distinguish between changes that are in line with the natural variation expected under that distribution and those that indicate some underlying change, like a difference in the average rate of occurrence.

B.41 If a change or difference is large enough to indicate an underlying change, it is said to be statistically significant (commonly abbreviated to significant). Statistical significance is expressed in terms of a confidence level, given as a percentage, and 95% is often used as a threshold to infer a statistically significant difference. This means that 19 times out of 20 the observed difference would be a result of a difference in the underlying distribution (like a change in the average rate), but there is still a 5% chance that the difference is the result of natural variation. If, for example, you have observed two collision rates, the probability that they arose from the same underlying distribution is known as the p-value. The smaller the p-value, the more confident you can be that observed differences are significant. It is important to note that the validity of the statistical test depends on the underlying assumptions holding (like collisions occurring independently of one another) and that a statistically significant difference (for example, between two averages) does not in itself indicate causality.

B.42 With large enough numbers, the Poisson distribution can be approximated by the more commonly known normal distribution. Probably the most common statistical test is for a difference in averages based on the normal distribution. And the normal approximation to an underlying Poisson distribution opens this up as a way to statistically test the difference between numbers (or rates) of counterfactual and after collisions on ALR motorways.

B.43 Highways England's *Overarching Safety Report* describes differences between the counterfactual and after collision or casualty *rates* as being statistically significant or not significant. This creates the impression that the tests were undertaken in this way, as a test of the difference in means, based on a normal distribution. However, based on historical precedent from how POPEs have been carried out, Highways England employed a different statistical test, a chi-squared goodness of fit test. This test is more typically used to test whether a distribution fits a prior expectation. For example, you could use the goodness of fit test to test whether road safety data follow a Poisson distribution and the difference in means test to test the difference between the average collision rates in two years or on two subsections of the network. Appendix 1 describes the two tests in the context of the safety impact of an ALR motorway.

B.44 In this case, where the goodness of fit test has one degree of freedom, and the difference in means test uses pooled variance, the two tests are equivalent. The result from the goodness of fit test is equal to the square of the result of the difference of means test and both tests produce the same p-value. This means that both tests provide an identical level of confidence in whether the difference between the counterfactual and after period is statistically significant.

B.45 Strictly speaking, the goodness of fit test can only reject the hypothesis that the underlying rate is the same in both periods; it cannot be used to infer any direction or magnitude. However, there are only two categories, so it is clear whether the after period is higher or lower than the counterfactual and the test is equivalent to a difference in means test, which is directional. So, this is not an issue in practice.

B.46 Given the equivalence, we do not have any fundamental concerns with the goodness of fit test in this context. Using the difference in means test would not improve the robustness of the results in general, but we feel it more directly addresses the question being asked. Why use a method that indirectly tests the difference between two rates, rather than one that does it directly? As it is more commonly used and understood, we believe that a difference in means test would be easier to carry out and review. It would also create a clearer, more transparent link between what is reported and the analysis that has been carried out (including the use of a rate-based counterfactual adjustment, which could have benefits when dealing with pandemic-era data. This is discussed in more detail in appendix 1).

B.47 There is one exception to this equivalence in Highways England's analysis. Highways England summed the expected collisions (or casualties) from the individual projects to form the expected collisions for ALR overall, instead of basing them on the overall collision rate and traffic levels. This did not affect the number of observed counterfactual or after collisions (or casualties) and the impact on the chi-squared test statistic (and associated p-value) would have had no bearing on the conclusions about statistical significance. However, that this discrepancy can arise from different ways of aggregating the data strengthens the case for adopting the more direct difference in means test in future analysis.

B.48 By pooling projects together, Highways England has created a dataset that meets the basic sample size requirements for the statistical tests. However, as we have already recommended, there would still be value in updating the analysis to include more recent data.

B.49 There are several other areas of development where **we recommend** Highways England should improve the robustness of its approach and/or how the results are interpreted in the future:

(a) Casualty numbers are not independent of one another, which negatively affect the robustness of the statistical tests on casualties and KSIs. We understand that Highways England is developing a method to account for this in confidence intervals and statistical tests. While this approach is being developed, Highways England should include a clear health warning around significance tests on casualty numbers or avoid carrying them out and focus on collisions.

(b) Linked to the point above, for transparency Highways England should set out its approach, and the assumptions that underpin it, more fully. This would not increase robustness in itself but would improve credibility and make it easier to understand the robustness of the evidence.

(c) Highways England's analysis of uncertainty in the *Overarching Safety Report* focused heavily on the tests of statistical significance. There is very little sensitivity testing to demonstrate the robustness of results to other sources of uncertainty. We think this should be strengthened in future and could cover areas such as the counterfactual (as discussed in the previous section) and testing the impact of individual projects on the overall ALR-level results.

(d) The high-level statistics by road type offer a comparison between ALR motorways and other motorways, including other forms of smart motorway. And the before and after analysis compares ALR motorways to what safety might have looked like on those roads if they had not been converted. Extending the scope of the before and after analysis to include DHS projects on a consistent basis would offer an extra level of comparison of the relative safety impact of different forms of smart motorway.

Evidence from other countries

B.50 Part of the scope of this review is to identify whether there is useful data from the international experience of operating similar schemes. This part of the review of the literature has focused mainly on before and after safety impact assessments and is distinct from the international comparisons in annex C. We have based our findings on a sample of the most relevant literature published on the internet. Most of the evidence we have used is drawn from academic literature.

B.51 We have limited our review to the conversion or use of the hard shoulder as a running lane. Consistent with our findings in annex C, all the studies we have found in other countries relate to hard shoulder running HSR rather than ALR. A full list of the papers we reviewed is given in appendix 2.

B.52 We have not attempted to chart the international development of HSR. However, based on the studies we have found, HSR has been deployed in a variety of different countries and is not especially new. We have gathered evidence from Germany, the Netherlands, France, the USA, and South Korea. Germany and the Netherlands appear to have been the first countries to adopt HSR. In Germany, HSR schemes date from the mid-1990s.

B.53 In all countries, HSR is deployed on a minority of the network and is reserved for busier stretches of highway. In Germany, HSR is regarded as an 'interim solution' to allow authorities to increase capacity until it is feasible to widen the carriageway, although in practice many HSR roads have operated for over two decades.

Evidence on the overall safety impacts of HSR

B.54 Whilst the experience of HSR or DHS is informative, we need to be cautious when drawing parallels with ALR. There are two aspects to consider. Firstly, intermittent use of the hard shoulder introduces risks that are not present with ALR (for example, potential confusion resulting from a hard shoulder being in place at some times and not at others). Secondly, the trade-offs between the loss of a hard shoulder and the safety benefits of improved traffic flows and/or reduced speeds will be different for HSR (where the hard shoulder is used as a live line during peak times) and ALR (where the hard shoulder is permanently converted to a live lane).

B.55 Moreover, both HSR and ALR schemes involve a package of measures to increase capacity and improve the smooth flow of traffic. Each of the schemes involve different packages of measures (for example, the introduction of HSR in combination with traffic management measures such as speed and lane controls) to improve operation or maintain safety. These features and the context within which the schemes are delivered will influence safety outcomes. In this review, we have not tried to unpick the influence of different design aspects and the literature is not sufficiently broad or detailed to do so. This is not to say that there are not benefits of international collaboration in respect of safety risks and design aspects (as we refer to in annex C).

B.56 The evidence on the safety impacts of HSR is mixed but, in the main, positive. Most of the studies we have reviewed conclude either that the introduction of HSR has delivered a reduction in collisions or that congestion relief has been achieved without any obvious or measurable impairment to safety. One of the studies we have looked at – *Safety Effects of Freeway Hard Shoulder Running* (South Korea) – found that some HSR sections were associated with an increase in collisions, whilst others showed a decrease.

B.57 Headline findings from before and after analysis from other countries is as follows:

- An analysis of seven HSR freeways in Germany found no evidence of an increase in crash frequencies or costs and concluded that HSR can improve road safety. (Paper 1).

- For a 2.3km section of urban motorway near Paris, researchers observed a reduction in collisions on the HSR section. However, this was partly counterbalanced by a migration of traffic density and collisions to the non-HSR section downstream. (Paper 4).

- An evaluation of an Active Traffic Management system in Virginia (USA) identified a statistically significant reduction in collisions following the introduction of traffic management measures and *dynamic* hard shoulder to sections of road where peak period HSR was already in place. (Paper 2).

- Significant improvements in safety have been observed where HSR has been implemented in the Netherlands. (Paper 8).

- A study concerning 22 sections of HSR in South Korea found evidence of an increase in collision rates. It concluded that caution should be exercised when designing and implementing HSR to reflect the trade-off between improvements in efficiency and a reduction in safety. (Paper 3).

Detailed findings

Impacts on more and less severe collisions

B.58 Some of the studies attempted to separate out the effects on more and less serious collisions although there are too few examples on which to draw general conclusions. The long-term assessment of HSR in Germany (Paper 1) looked at both the number of crashes and crash costs but found no evidence of a change in average crash costs which may indicate a shift towards more or fewer serious collisions. The study of HSR in South Korea (Paper 3), found that collisions involving injuries or fatalities increased on sections of HSR with 2(3) lanes. On sections with 4(5) lanes, the increase in collisions was due almost entirely to an increase in non-injury collision.

Upstream and downstream impacts

B.59 An interesting aspect of the safety analyses we have reviewed is the consideration given to potential safety implications for sections of highway upstream or downstream of the section of HSR itself. This does not appear to be an aspect of the safety debate considered in the UK. Highways England has stated that there are safety benefits from abstracting traffic from less safe, non-motorway roads. Highways England's POPE analyses, which generally cover a wider area than just the converted motorway links, provide an opportunity to investigate impacts on competing routes, as well adjacent sections of the SRN.

B.60 An assessment of a relatively short (2.3km) section of HSR near Paris (Paper 4) found a slight reduction in collisions on the section itself but found evidence of an increase in collisions on the section immediately downstream. The authors attributed the increase in accidents to the migration of traffic density from the HSR section to the downstream section. In contrast, the study of German HSR roads (Paper 1) found a relatively large (40%) and statistically significant reduction in collisions on three sections of road upstream of the HSR sections themselves. This, the authors conclude, is due to congestion relief on the upstream sections which itself is attributed to the implementation of HSR.

The determinants of safety impacts

B.61 Where safety benefits of HSR have been identified, these tend to be attributed to a reduction in traffic density or congestion afforded by the increase in capacity.

B.62 For example, a case study of the I66 Active Traffic Management system in Virginia, USA (Paper 2) also considered accident frequencies by type of collision. It concluded that there was a close relationship between the operational and safety benefits of HSR and found statistically significant reductions in crash types associated with congestion. The study found overall reductions in collisions on sections of the I66 with HSR but no such reduction on sections fitted with Active Traffic Management alone. It should be noted, however, that this study concerned the introduction of Active Traffic Management and dynamic hard shoulder on sections of highway that already utilised the hard shoulder during peak times (albeit for fixed hours).

B.63 The most comprehensive and detailed analysis amongst our sample is the long-term study of the safety impacts of HSR in Germany (Paper 1). This study was explicit in its attempt to unpick two competing factors; the safety benefits of reduced congestion and the reduced risk of rear-end collisions, versus the safety risks of the conversion of the hard shoulder to a running lane either because of the lack of a refuge area for broken down vehicles (when HSR is in operation) or due to vehicles attempting to use the hard shoulder when HSR is no longer in operation. The latter is an issue that applies to HSR but not ALR. Waleczek and Geistefeldt analysed data for seven HSR sections of between 2km and 19km in length constructed between 2001 and 2010. Based on the spatial distribution of crashes, they found no evidence of an increase in collisions involving stationary vehicles and concluded that the loss of a hard shoulder during peak hours did not affect safety. Although the frequency of crashes was higher during periods of HSR operation, this was entirely explained by differences in traffic flows during these periods. In respect of the safety benefits of reduced congestion, as noted above, the authors found no significant reduction in collisions on the HSR sections themselves but did find evidence of reduced rates of collisions on upstream sections afforded congestion relief.

B.64 Given the role that capacity and congestion relief might play in reducing the rate of collisions we also reviewed a US study specifically concerning the relationship between traffic density (traffic flows per lane) and safety. This study concerned urban freeways in Seattle and Minnesota. It finds a U-shaped relationship between traffic density and the rate of collisions. At very low levels of traffic the crash rate is very high. The authors attribute this, at least in part, to very high crash rates between midnight and 5 a.m. due to a mix of impaired and fatigued drivers with low traffic volumes (and therefore high speeds). Outside this period, at low to midrange levels of traffic density, crash rates remain low. Once a critical threshold combination of speed and density has been exceeded, the crash rate rises rapidly. It might be expected that, during periods of high traffic density, that a higher rate of less serious, lower speed collisions, but a lower rate of more serious, higher speed collisions would be observed. However, the Colorado study found that the U-shaped relationship between traffic density and crash rates held for both Property Damage Only (PDO) and Fatal and Injury (FI) crashes.

B.65 In an earlier study (Paper 9), Geistefeldt concluded that the traffic conditions during HSR operation are a crucial factor in determining safety impacts. His overall conclusion (based on data for a single 5km section of freeway) is that temporary hard shoulder running does not affect road safety. He suggests that the hard shoulder should be opened to moving vehicles when traffic levels reach a threshold at which congestion occurs and closed at other times. Extending this logic, he asserts that temporary hard shoulder running combined with dynamic traffic control systems (i.e. DHS) is favourable to permanent use of a hard shoulder as an additional lane.

B.66 This U-shaped relationship has been noted elsewhere, although we have made no attempt in the course of this work to consider how generalised this relationship is, at what levels of traffic an increase in collisions begin to kick-in, or whether it is borne out on the motorway network in the UK. However, we think that this would be a useful area of research to pursue.

B.67 Traffic density is just one factor that influences accident rates, but it does highlight the importance of the context in which HSR or ALR is implemented. For example, if a U-shaped relationship between traffic density holds, then this would suggest that (all things being equal) HSR or ALR is more likely to deliver overall safety benefits at locations and during periods of the day when traffic density is high. The

Stocktake attempts to draw general conclusions about the safety of different motorway types. An alternative perspective taken by some of the studies we have seen is to ask under what conditions ALR is likely to deliver safety benefits.

B.68 Two of the studies we have reviewed are predictive; they set out a methodology for predicting the impact of HSR implementation. One such study, which utilised data for four-lane freeways in Colorado, attempted to model the rate of collisions based on the level of service, itself a function of speed and traffic density. The authors used this model to estimate the reduction in collisions that would result from the increase in capacity afforded by HSR. The authors concluded that, given the density of traffic on the routes in question, the benefit of crash reduction is likely to outweigh the adverse effect of not having full hard shoulders. It should be noted, however, that this study used broad assumptions for the impact of the loss of the hard shoulder and did not separate out the implications for more serious of fatal collisions.

Observations on methods employed

B.69 We observed a range of different methods as part of our review. These range from relatively simple direct before and after comparisons of collision rates to more sophisticated statistical approaches. As noted, some of the studies focus simply on the overall number of collisions. The more instructive studies look at collisions of different types and severities.

B.70 There are two main areas where we have drawn comparison between the methods used in the POPE studies and *Overarching Safety Report*, and what might be considered best practice from the sample of international studies we reviewed. The approaches may be instructive for any further before and after analysis undertaken by Highways England in the future.

Defining the counterfactual

B.71 Most of the studies compare the post-implementation collision rates against a counterfactual. Consistent with Highways England's *Overarching Safety Report*, in all cases, the counterfactual is intended to represent a scenario in which no intervention was delivered (as opposed to a scenario of conventional widening).

B.72 The assessment of the section of HSR in France attempted to base the counterfactual on a single comparable site (in terms of traffic flows and lane width).

B.73 The more detailed studies use safety performance functions (SPFs). As referred to above, SPFs are simple regression models that predict the accident rate for different highway types (e.g. 3-lane freeway) based on a limited number of explanatory variables (such as traffic flows, average speeds, and the share of heavy goods traffic). SPFs are used widely in road safety analysis in the USA. The studies that used SPFs did not specifically reference the need to consider an underlying trend in collision rates (the driver of Highways England's approach to the counterfactual) although this might be less of an issue where the SPFs are updated regularly.

Tests of statistical significance

B.74 The more detailed studies we have looked at have attempted to establish whether differences in accident rates were statistically significant and employed a variety of methods.

B.75 An issue addressed by two of the studies we have reviewed is that of 'overdispersion'. This describes a situation where observations are highly clustered as can be the case with more severe collisions and fatalities particularly. This increases the risks that changes in the before and after accident rates are a product of the natural variation in collisions, or regression to the mean effects, rather than a change that can be attributed to the scheme in question. Papers 2 and 3 employ the 'Empirical Bayes' method. Under this approach, the SPFs (referenced above) include an overdispersion parameter which is used to weight the observed before and modelled counterfactual collision rates to control for overdispersion of the data.

Conclusions

B.76 In our view, before and after comparisons of safety performance (given sufficient data and if carried out effectively) provide the strongest form of evidence of the safety of ALR. This form of analysis is quite distinct from the high-level statistical comparisons as it goes some way towards isolating the net effect of converting a conventional motorway to ALR.

B.77 Highways England's approach to the before and after analysis was consistent with its standard in-house methodology for evaluating the safety impacts of major projects as part of the POPE process. This approach compares safety outcomes against a 'counterfactual' intended to represent a scenario in which the scheme had not been implemented. Highways England has then applied a test of whether the difference between the observed collision or casualty rates is statistically significant. In these regards, the approach is both in line with good practice and accords with guidance.

B.78 As compared to examples of similar analysis undertaken elsewhere, we have found the safety evaluation within the POPE methodology to be relatively basic. This offers advantages in terms of clarity and repeatability. The corollary to this is that there are also ways in which the analysis could be refined to add to its robustness. We have begun to explore some of these options in this review.

B.79 One of the limitations of the before and after evidence included in the *Stocktake* is that many of the ALR schemes had been in place for a relatively short period of time. This makes it more difficult to determine whether differences in the before and after collision and casualty rates are due to the ALR schemes, some other factor, or simply chance. The sample size requirements of statistical tests have been overcome by pooling the data for nine ALR schemes in Highways England's overarching analysis, this element of the analysis forms the basis of the conclusions in (paragraph 5.3 of) the *Stocktake*.

B.80 The reliance on pooled analysis requires an assumption that all the schemes are fully comparable; potentially masks scheme-level differences; and puts a lot of weight on a single year of data for the majority of the projects. Therefore, in the interests of both robustness and transparency, Highways England should update the before and after analysis contained within the *Overarching Safety Report* to include 2019 and 2020 data (if it is deemed sufficiently comparable) by the end of March 2022.

B.81 As part of this update, Highways England should implement a set of relatively minor improvements to its approach, including applying its latest method for counterfactual adjustments. Also, to allow comparisons to be drawn, **we recommend** that Highways England extends its scope to include equivalent analysis for DHS schemes.

B.82 Over a longer timescale, **we recommend** that Highways England undertakes a broader review of its approach to the safety evaluation of new highway schemes. This should be informed by a review of the full range of statistical methods used elsewhere to evaluate the safety impacts of highway schemes. Consideration should also be given to the timing of such analysis and whether the default approach of relying on one- and five-year POPE reports is appropriate.

Appendix 1 – Description of the statistical tests

The *Smart Motorway All Lane Running Overarching Safety Report* includes tests of statistical significance for four measures: personal injury collisions, KSI collisions, total casualties and KSI casualties. As noted, the test seeks to establish whether there is a statistically significant difference between the rate of collisions or casualties observed in the ‘after’ period, and the rate of collisions or casualties under the counterfactual.

For the counterfactual, Highways England applied an adjustment factor to the number of collisions or casualties observed in the ‘before’ period.

The chi-squared goodness of fit test

The standard form of the chi-squared goodness of fit test is set out below:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

Where:

χ^2 *is the chi square statistic used to generate the p − value*

O_i = *an observed count for bin i*

E_i = *an expected count for bin i, asserted by the null hypothesis*

n = *number of bins*

Highways England tested a null hypothesis of a constant collision or casualty rate in the before (counterfactual) and after periods against an alternative hypothesis that the rates are not constant. The way the calculation has been carried out in practice is as follows:

$$\chi^2 = \frac{(Obs'\ Before - Exp\ Before)^2}{Exp\ Before} + \frac{(Obs\ After - Exp\ After)^2}{Exp\ After}$$

Where:

$$Obs'Before = Obs\ Before \times CounterfactualAdjustment$$

$$Exp\ Before = TrafficBefore \times Combined\ Before\ (Ctfcl)\ \&\ After\ \ Rate$$

$$Exp\ After = TrafficAfter\ \times Combined\ (Cftcl.)\ Before\ \&\ After\ \ Rate$$

The ‘Combined Before (Ctfcl) & After Rate is calculated as follows:

$$Combined\ Before\ (Cftl.)\ \&\ After\ Rate = \frac{(Obs'\ Before + Obs\ After)}{Traffic\ Before + Traffic\ After}$$

The difference in means test

Alternatively, the difference between the before (counterfactual) and after rates could be tested with a difference in means test. This would also test a null hypothesis that the rates are equal against an alternative that they are not. This means undertaking the test with a pooled variance, which would take the form:

$Z = \frac{\hat{\alpha}_1 - \hat{\alpha}_2}{\sqrt{\hat{\alpha}}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$, where $\hat{\alpha} = \frac{n_1\hat{\alpha}_1 + n_2\hat{\alpha}_2}{n_1 + n_2}$

and:

$\hat{\alpha}_1 = before\ (counterfactual)\ collision\ or\ casualty\ rate$

$\hat{\alpha}_2 = after\ collision\ or\ casualty\ rate$

$n_i = sample\ size\ of\ period\ i\ (measured\ in\ traffic\ volume)$

The counterfactual adjustment

Because the goodness of fit test operates in absolute numbers of collisions (or casualties), in the overarching ALR analysis Highways England applied an adjustment based on the change in the absolute number of collisions (or casualties), rather than the change in the collision (or casualty) *rate* (for example per hundred million vehicle miles). The adjustment factors were based on the mid-points of the before and after periods.

As discussed in the main body of this annex, this is more appropriate when there is a clear trend over time, as is the case for total collisions (or casualties), and is less appropriate when there is a large degree of year-to-year volatility, as is the case for KSI collisions (or casualties).

The pandemic has substantially reduced traffic levels and, in all likelihood, the absolute number of collisions (and casualties). This could have implications for how Highways England applies its counterfactual adjustments. For example, if an after period is from 2018-2020, the counterfactual adjustment would be based on 2019 data and would be unaffected by the pandemic. But the absolute number of collisions or casualties in 2020 would likely be substantially lower, potentially artificially improving the safety performance of the scheme. Assuming that collision (and casualty) *rates* are less likely to have differed substantially during the pandemic, rate-based counterfactual adjustments would be preferable to those based on absolute numbers.

2020 data should only be included in updated analysis if it is deemed sufficiently comparable to the rest of the time series. An advantage of the difference in means method is that a rate-based counterfactual adjustment could easily be used.

Appendix 2 – International literature review

(1) Helen Waleczek and Justin Geistefeldt (2021). **Long Term Safety Analysis of Hard Shoulder Running on Freeways in Germany.** (Transport Research Record).

(2) Nancy Dutta, Michael D. Fontain, Richard Atta Boateng (2018). **Evaluation of the Impact of the I-66 Active Traffic Management System: Phase II.** (Virginia Transport Research Council).

(3) Jaisung Choi,Richard Tay, Sangyoup Kim, Seungwon Jeong, Jeongmin Kim and Tae-Young Heo (2019). **Safety Effects of Freeway Hard Shoulder Running**. (Applied Sciences).

(4) M. Aron, S. Cohen, R. Seidowsky (2013). **Safety impact of using the hard shoulder during congested traffic: the case of a managed lane operation on a French urban motorway.** Transportation Research Part C: Emerging Technologies.

(5) Marco Guerrieri, Raffaele Mauro (2016). **Capacity and safety analysis of hard-shoulder running (HSR). A motorway case study.** Transportation Research Part A: Policy and Practice.

(6) Jake Kononov, Steven Hersey, David Reeves, and Bryan K. Allery (2012). **Relationship Between Freeway Flow Parameters and Safety and Its Implications for Hard Shoulder Running.** Transport Research Record.

(7) Douglas W. Harwood, Karin M. Bauer, Ingrid B. Potts (2013). **Development of Relationships between Safety and Congestion for Urban Freeways.** Transportation Research Record.

(8) Conference of European Directors of Roads (CEDR), Traffic and Network Management Working Group (2018). **Hard Shoulder Running Fact Sheet.**

(9) Justin Geistefelt (2012). **Operational Experience with Temporary Hard Shoulder Running in Germany.** Transportation Research Record.

This publication is available at orr.gov.uk

Any enquiries regarding this publication should be sent to us at orr.gov.uk/contact-us