Reported Road casualties – Data accessibility project: Final Report

Statistical Disclosure Control
Strategy and Standards Directorate
Office for National Statistics

Summary

The Department for Transport publish annual tables of road traffic accident statistics using descriptive variables such as road type, accident severity and vehicle type. Some tables show Contributory Factors as well; subjective information collected by a police officer attending the accident which can help explain how the accident happened and possibly also indicates blame. A formal statistical disclosure control policy for the release of these data is required.

This paper gives recommendations and defines a structure for the release of tables containing Contributory Factors. Key characteristics of the data for a range of tables are examined. These include intruder scenarios (how disclosure breaches may occur) alongside risk measures (risk is related to table detail, geographical level etc) and result in a range of risk categories for these tables. The management of this risk is also described in terms of how protection can ensure release. Disclosure scenarios and protection methods are described and there is a brief discussion on releasing microdata to researchers.

1. Introduction

Reported road casualties for the previous year are published each September in an Annual Report, following an initial publication in June. A large number of tables of counts and percentages are released showing information such as reported road accidents under specific road conditions and the numbers of casualties and vehicles involved. Typical variables which can be included in these outputs are age of driver/ casualty, road user type and severity of accident. These variables will be discussed in more detail in the next section.

This report concentrates on a subset of the tables published in the annual report, those concerning Contributory factors (CF). These are shown in detail in Appendix 2. More than one CF (up to 6) can be issued for each accident and are determined by the police officer attending the scene of the accident in a 'STATS 19' report. These can aid the investigation but also could potentially suggest a legal liability and thus cause distress to an individual if identified.

The general aim is to review and update disclosure control policy for road casualty statistics and in particular decide on a policy for outputs involving CFs.

The outputs are to include specific recommendations and guidance on the data that could be published and a checklist for assessing future requests.

A broader question to be answered is to investigate approaches to balancing the demands of FOI and the transparency agenda against data protection issues.

2. Defining the variables and the problem

The main aim of the project is to determine a policy for the release of the CF data which is an extensive resource but currently underused. There are 77 CFs from which a police officer is able to select up to 6 as possible influencers. These are shown in Appendix 2 along with the 9 main categories in which they fit. There is much information here which could be used to give a detailed breakdown of causes of accidents and how they relate to the factual variables which are collected by Department for Transport (DfT) (see Appendix 1 for more detail). These variables define the accident in relation to physical conditions (road, weather etc.), vehicle characteristics and casualty detail.

It needs to be emphasised that CFs are largely subjective and indicate from the perspective of a police officer the actions and failures which lead to the accident. There will also be considerable variability between Police Officers (some more knowledgeable about traffic conditions etc) and Police Forces (some forces give moiré training relating to the completion of the accident form and expect CFs to be thought about carefully).

This will have an effect on the quality of the outputs and will also increase the uncertainty of any identification of individuals being made from the data. Uncertainty will also be increased by the fact that not all accidents are included in outputs using CF data. Only accidents attended by the police and where they reported at least one CF are included. In 2011 78% of accidents met this criterion.

Tables RAS50001 to RAS50014 released in September 2012 from 2011 data involve CFs. These are generally counts (and percentages) of each CF for **one** particular reported factual variable such as Severity of Accident or Road Class at National Level with CF counts for each Country and Region also published. Counts can be based on accident, casualty or vehicle. The full list of published tables and the relevant variables is:

- RAS50001: Severity (Fatal, serious, slight)
- RAS50002: Counts for 10 most frequently reported CFs
- RAS50003: Road type (Motorway, A, B, other)
- RAS50004: Counts for 10 most frequently reported CFs involving pedestrians
- RAS50005: Vehicle type (Pedal cycle, motor cycle, car, bus or coach, LGV, HGV)
- RAS50006: The most frequent pairs of contributory factors assigned to the same vehicle or pedestrian casualty
- RAS50007: Casualties in reported accidents by severity (Fatal, serious, slight)
- RAS50008: Number of accidents and resulting casualties where 2 speed factors were reported
- RAS50009: Number of vehicles with 2 speed contributory factors by selected vehicle type
- RAS50010: Percentage of vehicles with selected contributory factors which also had a speed factor reported
- RAS50011: Reported accidents and vehicles included in the contributory factor analysis:
- Proportion of accidents and vehicles with at least 1 contributory factory and with a police officer attending
- RAS50012: All contributory factors by England, Wales and Scotland
- RAS50013: Contributory factors: Reported accidents by Region

- RAS50014: Two vehicle accidents in which a driver or rider had "failed to look properly" as a contributory factor

Only a limited number of the factual variables have been utilised here. Much more could be investigated by looking at variables such as age of driver/casualty and speed limit. This is where the issue of statistical disclosure control will need to be considered.

There will be an element of risk if tables are published with increased sparsity and with low frequencies. Questions to be asked are:

- Could individuals in these tables be identified (including self identification)?
- Can attributes relating to individuals be discovered?
- What information in the public domain could be used along with these published tables to lead to possible disclosure?

3. Defining Risk

The previous section ended by discussing the possibility that more detailed tables could be considered risky, i.e. potentially disclosive. This can be examined further by thinking about intruder scenarios, how might an attacker (intruder) discover something from the data?

The possible intruder scenarios to consider are shown here.

- Motivated attacker. An individual who studies the tables and notices low frequencies. Uses additional sources of publicly available information such as newspapers, court records, on line chatter to attempt to identify an individual. The key point here is that sources ought to be likely reasonably to be used to identify an individual and reveal information about them. Therefore all local sources need not be taken into account, only information likely to be available to an intruder.
- Self Identification. A count of 1 in a table could enable somebody to spot themselves. How much of a problem is this? Could it be traumatic? Maybe only for the most sensitive variables. These could require more protection.
- Nosy neighbour. Somebody who sees an accident or the results of an accident and uses the tables to find out more, especially when noticing low frequencies. Maybe less likely to use additional data sources than a motivated attacker.

Any form of identification in a table with a large count is not likely to be as serious as the absence of uniqueness is protection against disclosiveness. This lack of rarity will discourage an attacker from trying to find out anything about an individual in a cell and will also alleviate concerns relating to self identification.

In general, outputs can be placed into one of three broad risk categories, defined in terms of the likelihood of an attempt to identify individuals, and the impact of any identification.

These definitions are from the GSS disclosure control policy for tables produced from administrative sources.

Risk measures

- Low Risk: High level of aggregation with only limited tables produced from the database. Little protection required apart from good table design. Care needs to be taken where rows and columns are dominated by zeros and where a marginal total is 1 or 2. In most cases the table can be published as it stands.
- Medium risk: Smaller populations than above, such as lower level of geography. Tables may be linked. Cells of size 1 and 2 maybe considered unsafe.
- High Risk: The impact of a successful identification will be great, data are especially sensitive. Data are likely to be produced for small populations. All cells of size 1 to 4 are considered unsafe.

The scenarios below define risk levels for a range of tables with low counts and give suggestions for the protecting the table, if necessary.

4. Scenarios/Case Studies

A number of tabulations from the 2011 data are described in the scenarios in this section. These show output combinations with low frequencies where a CF assigned by the police officer is tabulated. These tabulations are potentially disclosive and each is discussed individually.
Prior to looking at individual scenarios a couple of general points are raised. These are User requirements (who will use the data and why) and understanding the key characteristics of the data (level of risk of disclosure, quality of the data).
For each scenario possibilities of disclosure will be described and discussed. Suitable disclosure control methods which assist in managing the risk will be summarised for all scenarios as will legal aspects and details of implementation.

The points to be discussed for the scenarios are

- What are the users' requirements? Why would a table such as this be of interest?
- Understand the key characteristics of the data. Quality (CFs are subjective). Are the data relating to accidents, casualties or vehicles?)Which variables in the table are riskier?
- Circumstances where disclosure is likely to occur. Relate this to the intruder scenarios. Self identification? What might a motivated intruder (or nosy neighbour) be looking for? What attributes might they want to discover? This will lead on to the different risk categories. Can tables at a particular output level (national, regional etc) be associated with the same risk level? What other issue need to be considered? Data from other sources or linking tables?
- Does any disclosure risk identified represent a breach of public trust, law or policy? Are personal data to be released? These are National Statistics. What does this mean legally?
- Select appropriate disclosure control methods to manage the risk. Recoding and table design will probably be the most appropriate. Mention the alternatives briefly.
- Implement and disseminate. If disclosure control has been applied clearly state this is the case. Let the user know if any changes are made to the SDC policy.

The CF variables have been rated for visibility and sensitivity, defined as follows:

- Visibility: How obvious would this be to a passer -by / neighbour etc

- Sensitivity: How much damage will this cause to the motorist if this CF can be connected with an accident, a vehicle or a casualty. The more sensitive a variable the greater the impact a disclosure will be.

These are attempts to highlight the factors for which blame can be apportioned and shown to be apportioned.

When analysing the data, separate tables were created for the CFs defined as **high** for both visibility and sensitivity and those defined as medium or low for either/or visibility and sensitivity (**other**). These are shown in appendix 2.
This is a subjective attempt to categorise an already subjective variable. If this proves to be a worthwhile approach the outputs with 'other' CFs would be defined as lower risk tables. Each scenario will consider outputs where 'High' CFs are tabulated.

Users' requirements: Users of transport statistics in general and road traffic accident statistics in particular are numerous (National and local government, pressure groups, members of the public). These data are easy for the public to relate to and outputs showing clearly the frequency of road accidents with as much detail (both factual and subjective) will be heavily utilised. Therefore the aim is to produce straightforward tables with as much detail as possible. The onus of these case studies is to justify not releasing data with the initial hypothesis being that as much as possible can be released.

Understand the key characteristics of the data: There are particular aspects of the data which will have an effect on the level and nature of the risk of disclosure. Some CF variables are more sensitive than others and can indicate blame. As described above, these have been assessed for visibility and sensitivity which could enable the factors where blame can be linked to an individual to be treated differently. Other characteristics of these data concern quality. The CF data are both subjective (some police officers and police forces will place greater emphasis on these than others) and not complete (over 20% of accidents have no CFs attached and are not reported in these tables). Any intruder would not necessarily know that an accident of which they had knowledge would not be included in these tables due to a lack of CFs. As road accident data are usually event based rather than residence based it is more difficult for an intruder to know the population at risk. It is harder to estimate who might have been travelling in a particular area during a specified time period than the number of people living in a particular location at any one time.

Factual variables also ought to be considered. A table with detailed factual variables could assist an attacker in identifying an individual, accident or vehicle and thus allow particular CFs to be associated with that individual, accident or vehicle. Some factual variables can be recoded to reduce this effect, Age can be recoded into bands of 5 or 10 years, the number of road user types can be reduced and speed limits can be banded.

Scenario 1
National Level tables with a single factual variable by CF. A number of tables are already published at this level.

Table 1

Table	Factual variable	Contributory Factor	Count
RAS50001	Accident Severity = Fatal	Vehicle door opened or closed negligently	1
Not published	Speed limit = 20mph	Disobeyed Double white lines	1
Not published	Speed limit = 50mph	Driving to slow for conditions or slow vehicle	2

Circumstances where disclosure is likely to occur

Cells with low counts are potentially disclosive. A table of Vehicle type * High CF has 34 cells with a count of 1 out of 522 cells in total. Further examples are shown in table 1.

Self identification

This could be an issue but there would be uncertainty regarding the role played by the police officer. A surviving individual would know they had been involved in a fatal accident as in the example from RA50001in tale 1 but how likely is it that they would they know the CF(s) reported by a police officer, if any had been reported. A driver ought to be aware of deficiencies with their car or the road but there would be a strong element of doubt in any identification of oneself at this high geographical level.

Motivated intruder / nosy neighbour

It is unlikely that any identification by an individual not involved in the accident such as a motivated intruder could occur. An intruder who has knowledge of an accident through a newspaper report (or through viewing it occur, however unlikely this may be) could make an educated guess about the associated CF(s) and may try and use other information in the public domain but there would be considerable uncertainty about whether discovered the correct accident.

For some of the more visible CFs (such as Defective traffic signals) it might be possible to judge which CFs have been included but there would be an element of doubt (Has the officer included this factor? Was the accident ever reported?)

At a national level, identification disclosure by an intruder would be unlikely as they would have to second guess the police officer's assessment. Even witnessing the accident except in the most unusual circumstances (animal in carriageway which is not a CF with both high visibility and sensitivity) would make identification difficult. The only attribute which could be disclosed would be the CF (assuming the factual variable is unique or rare in the table) and has been discussed previously the probability of an attacker associating the correct CF to an accident is low.

It is not easy to see how the low counts at a national level could be traced back to an individual by an intruder as the CF data would not explicitly be published elsewhere (although a cause identifiable as a CF could be mentioned in an inquest if there was a fatality). Also tracing data published at vehicle, casualty or accident level back to an individual would not be easy.

There is a utility issue with some of tables produced with a single factual variable and a single CF. It is doubtful how much can be gained from knowing an accident occurred where the speed limit was 20 mph and where a CF was 'Disobeyed double white lines'.

Risk category: Low risk – As long as the table is well designed there should be no problems with releasing the table. Any cells with counts of 1 or 2 should be looked at closely as identification of oneself or the other member of the cell is possible, although unlikely. If a factual variable category is unique or rare in a table (i.e. the marginal total is 1 or 2) care should be taken to ensure any visible and sensitive CF cannot be identified. However this is improbable at a national level.

Scenario 2

National level tables with pairs of factual variables by CF. There are more variables in these outputs than are currently published

Table 2

Table	Factual variable 1	Factual variable 2	Contributory Factor	Count
Not published	Accident Severity = Fatal	Vehicle type = Car	Vehicle door opened or closed negligently	1
Not published	Accident Severity = Fatal	Vehicle type = Pedal cycle	Junction Overshoot	2

Circumstances where disclosure is likely to occur

A table of High CF * Accident Severity *Vehicle Type has 2700 cells of which 45 are 1s.

Self Identification

Self identification could occur, and with an additional variable in the table and more low counts it may be more likely than for scenario 1. A surviving individual involved in a fatal accident (possibly the driver or passenger) involving a pedal cyclist who overshot a junction would know they were one of only 2 accidents assuming the relevant CF was recorded by the police officer. This strong element of doubt is a still a major influence on whether a cell such as this in a table can be considered disclosive.

Motivated intruder / nosy neighbour

A motivated intruder will have more information to use than for scenario 1. For example an individual with a particular interest in cycle accidents would see a cell count of 2 and investigate further, maybe to find out if an accident occurred at a particularly dangerous junction. This would be difficult at a national level however as attempting to search through newspaper reports and online resources of all cycle deaths in order to find a reference to a junction overshoot (even assuming the fatality was a cyclist and the junction overshoot was mentioned in any report) may be regarded as greater than reasonable effort.

The same would apply to an inquisitive bystander who may have witnessed an accident and attempted to search for further details.

Attribute disclosure would be possible if through a combination of unique CF and factual variable, the category of the other factual variable could be discovered. In table 2 an intruder might hear anecdotally of a car accident where the door of the vehicle was opened inappropriately and use this information to discover a fatality resulted. This is slightly implausible (certainty about a CF is unlikely) but probably ought to be taken into account.

Risk category: Low risk – This is not too dissimilar to scenario 1. A well designed table can be released following some checks. Any cells with counts of 1 or 2 should be looked at closely as identification of oneself or the other member of the cell is possible. Unique or rare combination of factual variables) care should be taken to ensure any visible and sensitive CF cannot be identified.

Scenario 3a

Country level tables with a single factual variable by CF

Table 3a

Table	Country	Factual variable	Contributory Factor	Count
Not published	Wales	Accident Severity = Serious	Disobeyed pedestrian crossing	1
Not published	Scotland	Accident Severity = Fatal	Disobeyed automatic traffic signal	1
Not published	Wales	Casualty seriously injured	Driving too slow for conditions	2

Circumstances where disclosure is likely to occur

At this lower geographical level the population at risk is obviously smaller that at a national level and the numbers in each cell will on average be smaller. Identification of an individual is potentially easier.

Self Identification

Self identification will be a little easier than in any national table and by extension easier in Wales and Scotland than England assuming the individual involved was aware of which country they were traveling in a the time of the accident. A driver in Scotland involved in a fatal accident might suspect the death of a pedestrian would have been assigned to the CF 'Disobeyed automatic traffic signal' but would not know for certain.

Motivated intruder / nosy neighbour

A motivated intruder would be able to narrow down any research to the particular country. It would still be difficult however to identify an individual or associated attributes. In table 3a an intruder may know somebody who had been seriously injured in a road accident in Wales. It would not be easy from knowing this to identify them as having disobeyed a pedestrian crossing as there will have been many other serious road accidents in Wales during the year. Due to the limited nature of these tables little or no attribute disclosure would be possible.

Attribute disclosure would be possible if the category of factual variable was unique or rare at country level thus allowing an intruder the possibility of finding out the CF. This appears to be unlikely.

Risk category: Low risk – As with scenarios 1 and 2 there would need to be considerable effort and guesswork to identify an individual with any certainty. Any cells with counts of 1 or 2 should be checked to confirm the unlikelihood of identification. Unique or rare factual variables within a country care should be taken to ensure any visible and sensitive CF cannot be identified.

Scenario 3b

Country level tables with pairs of factual variables by CF

Table 3b

Table	Country	Factual variable 1	Factual variable 2	Contributory Factor	Count
Not published	Scotland	Accident Severity = Serious	Age = 4	Loss of control	1
Not published	England	Accident Severity = Fatal	Age = 71	Travelling too fast for conditions	1
Not published	Scotland	Accident Severity = Fatal	Age = 15	Travelling too fast for conditions	2

Circumstances where disclosure is likely to occur

This scenario shows cell counts at country level for 2 factual variables and a CF. One of the factual variables in the all the examples in table 3b is age, which if known will increase the likelihood of identification considerably.

Self identification

There is some potential for self identification with the increased detail in table 3b, especially if one of the factual variables is individual age. Knowing one's own age and another factual variable along with the country could lead to self identification and possibly to identification of a CF. If this indicates blame (unjustified according to the driver) there could be consequences.

Motivated intruder / nosy neighbour

Examples such as shown in table 3b could allow motivated intruder to identify somebody and discover an associated attribute. Knowing the age of an individual along with the knowledge they were involved in an accident (serious or fatal) could enable a possible contributory factor to be inferred with a considerable level of confidence.

The parents of the 4 year old (if they were not involved in the accident) would be able to see the counts for each CF and with additional knowledge of the accident decide which CF referred to 'their' accident. If they thought the CF was loss of control it would allow them to blame the driver for the accident.

Age can be recoded grouped although some low counts particularly for Wales and Scotland still remain. For example If Age is recoded into the following categories
0-5, 6-15, 16-24, 25-44, 45-64, 65+
There are
4 serious accidents in Scotland involving ages 0-5 for 'Loss of Control'
5 fatalities in Scotland for ages 6-15 for 'Travelling too fast for conditions'
32 fatalities in England for ages 65+ for 'Travelling too fast for conditions'
But
1 fatality in Wales ages 0-5 for 'Travelling too fast for conditions'

If a factual variable is not as specific as Age it would be harder (although not impossible) to an attacker to identify an individual and possibly find out an associated CF.

Risk category: Medium risk: There is more risk attached to outputs with this level of detail with some due to an identifiable variable such as age being included. Maybe individual ages should not be used in the outputs but even when recoded there is a possibility of disclosure. Low counts in a table could encourage intruders to investigate further, therefore all cells of size 1 and 2 should be considered at risk. Large numbers of zeros in a row or column should also be looked at prior to publication to ensure all accidents do not occur in the same subgroup.
It could be worth considering whether these restriction need apply to the CFs defined as 'Other' as releasing these details would not lead to any blame being given for the accident. However as these data could be disclosive, considerable care would be required.

Scenario 4a
Region level tables with single factual variable by CF
Table 4a

Table	Region	Factual variable	Contributory Factor	Count
Not published	North East	Accident Severity = Fatal	Failed to signal or misleading signal	1
Not published	North East	Accident Severity = Serious	Driver using mobile phone	2

Circumstances where disclosure is likely to occur: The level of geography is much lower than country and therefore self identification should be more straightforward. There would be doubt about the region in which one was travelling.
Self identification
Examples in table 4a show how self identification can occur. A driver involved in a serious accident (to them self, a passenger or a pedestrian) in the North East whilst using a mobile phone would be able to recognize themselves. It is assumed this CF would have been chosen because it was reported to the police officer either by the driver or a witness. In both cases the driver would know mobile phone use has been associated with them.
Motivated intruder / nosy neighbour
A motivated intruder would also be able to narrow down any research to a relatively small location. It would still be difficult however to identify an individual or associated attributes. In the table 4a an intruder may know somebody who had been seriously injured in a road accident in the North East but it would be difficult for them to determine a CF as the number of accidents would still be reasonably large over the year.

Risk category: Low risk – This is not too dissimilar to scenario 2a. There would need to be considerable effort and guesswork to identify an individual with any certainty. Any cells with counts of 1 or 2 should be checked to confirm the unlikelihood of identification. Unique or rare factual variables within a country care should be taken to ensure any visible and sensitive CF cannot be identified

Scenario 4b
Region level tables with pairs of factual variables by CF

Table 4b

Table	Region	Factual variable 1	Factual variable 2	Contributory Factor	Count
Not published	North East	Accident Severity = Fatal	Age = 9	Exceeding Speed Limit	1
Not published	North East	Casualty = Serious injury	Age =40	Loss of Control	2

Circumstances where disclosure is likely to occur

This scenario shows cell counts at region level for two factual variables and a CF. As with the data at Country level, age is one of the factual variables leading to some disclosure issues.

Self identification

Self identification from table 4b would not be difficult. For specific combinations of factual variables, especially if one is individual age, a person may make assumptions about what CF has been attributed to them.

Motivated intruder / nosy neighbour

Using the information in table 4b, an intruder would be able to use a pair of factual variables such as knowing a 40 year old was seriously injured and link this to a CF. Other 40 year olds may have been injured in accidents during the year but combing this information with local news items may enable them to associate a pejorative CF to an individual.

Recoding age increases the counts in some but not all categories. For example there are 42 serious injuries ages 25-55 with CF 'Loss of Control' but 1 person aged 6-15 is a fatal casualty for CF 'Exceeding Speed Limit'.

Risk category: Medium risk: Similar to scenario 3b. There is greater disclosure risk attached to outputs with this level of detail often because an identifiable variable such as age is included. Maybe individual ages should not be used in the outputs but even when recoded there is a possibility of disclosure. Low counts in a table could encourage intruders to investigate further. All cells of size 1 and 2 should be examined. Large numbers of zeros in a row or column should also be looked at prior to publication to ensure all accidents do not occur in the same subgroup.

It could be worth considering whether these restriction need apply to the CFs defined as 'Other' as releasing these details would not lead to any blame being given for the accident. However as these data could be disclosive considerable care would be required.

Scenario 5a

LA level tables with single factual variable by CF

Table 5a

Table	LA	Factual variable	Contributory Factor	Count
Not published	Southwark	Accident Severity = Fatal	Swerved	1
Not published	Southwark	Accident Severity = Serious	Driver Using Mobile Phone	1

Circumstances where disclosure is likely to occur
At Local Authority level there is an increased likelihood of disclosure due to the smaller populations under consideration.

Self identification
From table 5a self identification can be seen. A driver involved in a serious accident in this LA whilst using a mobile phone would be able to recognize themselves. It is assumed this CF would have been chosen because it was reported to the police officer either by the driver or a witness. In both cases the driver would know mobile phone use has been associated with them.

Motivated intruder / nosy neighbour
A motivated intruder or nosy neighbour may be able to narrow their search down to a relatively small location. LAs vary considerably in size (both population and geography) so the possibility of identifying an individual will vary across the country. Using the level of detail as shown in table 5a an intruder would find it difficult to associate these data with a particular individual unless they had additional information such as a news report or court records stating for example that an accident in this LA was caused by mobile phone use or the factual variable was unique in the LA.

Risk category: High Risk: The detail of information which would be published here is not high but the population at risk is smaller than for other scenarios (even allowing for the fact that the output will consider the location of the accident not the LA in which the individual lives). There is also the perception issue. An attacker may consider trying to find out about individuals in a table at LA level simply because it looks a relatively soft target. Some of the information (especially CFs) that they are able to link to an individual could be correct and cause distress.
All cells of size 1 to 4 are considered unsafe. Large numbers of zeros in a row or column should also be looked at prior to publication to ensure all accidents do not occur in the same subgroup. If there is a need for detailed releases at LA level the CFs defined as 'Other' could be possibly treated as Medium risk

Scenario 5b
LA level tables with pairs of factual variables by CF.
Tables are very sparse at this level

Table 5b

Table	Region	Factual variable 1	Factual variable 2	Contributory Factor	Count
Not published	Southwark	Accident Severity = Serious	Age = 55	Disobeyed automatic traffic signal	1

The count is still 1 when age is grouped 45-64

Circumstances where disclosure is likely to occur:
As with scenario 5a there is an increased likelihood of disclosure due to the smaller populations under consideration. There is an additional variable with these releases with an increased possibility of disclosure.

Self identification
As under previous scenarios, self identification from table 5b would not be difficult and an individual aged 55 involved in a serious accident would probably be able to determine which of the CFs involved them.

Motivated intruder / nosy neighbour
An intruder could use a sparse table such as this to find out CFs relating to accidents in which acquaintances whose age they know, were involved.

Risk category: High Risk: There is more detail in these outputs than those discussed in scenario 5a. Any identification of an individual could cause great distress.
All cells of size 1 to 4 are considered unsafe. Large numbers of zeros in a row or column should also be looked at prior to publication to ensure all accidents do not occur in the same subgroup.

Other disclosure scenarios
A row or column in a table where all the counts fall into a single category could be disclosive. By knowing a colleague was a casualty in a fatal accident an intruder would also know the accompanying CF. If this scenario (known as group disclosure) is a considered to be a problem for specific tables then protection methods described in the Managing Risk section below should be employed.

CF ' Loss of Control'	Count
Casualty	
Fatal	4
Serious	0
Slight	0
Total	**4**

More detailed tables can also be requested, which may include more than one CF. Tables published with pairs of CFs would enable an attacker to find out more information about an accident as perceived by a police officer and maybe relate it to an individual if the counts are low. A similar process to that followed in the above scenarios can be employed for these more detailed outputs. If there are many variables in a table there is greater potential for the table to be disclosive. However many factors need to be considered such as the visibility and sensitivity of each variable and the level of geography at which the table is to be released.

Breath test data
Information is collected if a breath test is carried out. This is a factual variable but is potentially more sensitive as it can be related to a CF such as 'Impaired by alcohol' or 'Impaired by drugs'. The most disclosive scenario will be if a passenger involved in an accident knows the driver has either taken a breath test (but does not know the outcome) or that the driver has refused a breath test. If this information is published at a low geographical level they passenger could relate knowledge of the accident to details of the breath test and CFs.
Consider a theoretical table published at LA level where Accident Severity = Fatal, Blood Test = refused and CF = Impaired by alcohol. The passenger (or a relation/friend) may be able to identify themselves as being involved in this crash and confirm their suspicions that the driver was drunk.
In general outputs involving breath tests should be regarded as high risk and low cell counts should be protected.

Managing risk – If there is potential risk in an output any method of alleviating the risk will cause a reduction in utility. In protecting a table the requirements of the user(s) need to be kept in mind. If they are interested in particular variable combinations these should be maintained if possible.

If it is necessary to protect low cell counts there are a number of approaches. The recommended ones are discussed here.

- Table Redesign. Categories for certain variables can be combined to reduce table detail (and consequently utility). How this is carried out will depend on the structure of the table being produced and also on the requirements of the user. One possibility is to publish CFs at a high level only. The number of categories for this variable will be reduced from 78 to 9 thus leading to greater frequencies in most if not all cells. If the CF detail is necessary then a factual variable can be recoded. For example the speed limit can be shown in bands or age can be recategorised as described in scenario 3b. Any table redesign will rely on knowledge of both the data and the requirements of the user. The aim is to produce a table without low counts which is also relevant to the user. There will only be specific Age groups which will be useful and are the same as in other publications. Likewise combing fatal and serious accident severity makes more sense than combining serious and slight severity.
- Rounding. Here the categories are left untouched but each cell is rounded to a specified base, typically 3 or 5. There are different methods of rounding (simple, random, controlled) but the outcome is that each cell will be a multiple of the rounding base or 0. This approach can be used when the complete table is required but users will need to be aware of the resulting loss of information. This will be more noticeable when there are a large number of cells with small frequencies.
- Suppression. Low counts are suppressed and replaced with a symbol. Other cells will also need to be suppressed to avoid disclosure by differencing. For a small table the resulting lowering of utility can be high.

Identifying and managing disclosure risk – a summary

Table 6 summarises the above scenarios, including advice on identifying potential disclosure risks and how to mitigate these risks. In addition to the scenarios described previously table 6 will consider possible outcomes where 'Other' CFs are tabulated.

Table 6: A summary of disclosure scenarios and protection methods

Scenario	Geography level	No of factual variables	Contributory factor visibility / sensitivity rating	Risk category of table	Action	Protection (if required)
1	National	1	High	Low	Check marginal totals.	Recode the CFs to the higher level and/or recode the factual variable if possible
1	National	1	Other	Low	OK to release	-
2	National	2	High	Low	Check	Recode the CFs to the

					marginal totals	higher level or recode the factual variable if possible.
2	National	12	Other	Low	OK to release	-
3a	Country	1	High	Low	Check marginal totals	Recode the CFs to the higher level and/or recode the factual variable if possible
3a	Country	1	Other	Low	OK to release	-
3b	Country	2	High	Medium	Cells of size 1and 2 may be unsafe	Recoding as for low risk although rounding or suppression could be an option if low counts remain
3b	Country	2	Other	Low	Check marginal totals.	Recode the CFS and/or any suitable factual variable
4a	Region	1	High	Low	Check marginal totals	Recode the CFs to the higher level and/or recode the factual variable if possible
4a	Region	1	Other	Low	OK to release	-
4b	Region	2	High	Medium	Cells of size 1and 2 may be unsafe	Recoding as for low risk although rounding or suppression could be an option if low counts remain
4b	Region	2	Other	Low	Check marginal totals	Recode the CFs to the higher level and/or recode the factual variable if possible
5a	LA	1	High	High	Cells of size 1 to 4 are considered unsafe	The first step will be to recode the CF and Factual variable. If the table is not too sparse then rounding may be appropriate. Suppress small cells if necessary
5a	LA	1	Other	Medium	Cells of size 1and 2 may be unsafe	Recoding may suffice but rounding and suppression are options
5b	LA	2	High	High	Cells of size 1 to 4 are considered	Recode the CF and Factual variables. If the table is not too sparse

					unsafe	then rounding may be appropriate. Suppress small cells if necessary
5b	LA	2	Other	Medium	Cells of size 1and 2 may be unsafe	Recoding may suffice but rounding and suppression are options

If the table requires protection use the approaches described above. The relevant Contributory Factors can be recoded to the higher level. If a factual variable is to be recoded, age can be banded into 10 year ranges for example. Road user type and vehicle type can be coarsened. The method of recoding would depend on the requirements of the data user.

Rounding or suppressions should be considered if recoding is not feasible. Please note the limitations of these approaches with respect to information loss.

Example

Table 6 has indicated table risk categories, possible actions to take and protection methods that can be taken. Table 7 explores three potential risks using synthetic data alongside actual variables. The advice is not over specific as additional factors such as expert knowledge of the data is also important

Table 7: Individual Age by Accident Severity for specified CFs

		Region A										
		Age										
First CF	**Accident Severity**	**31**	**32**	**33**	**34**	**35**	**36**	**37**	**38**	**39**	**40**	**Total**
Exceeding Speed limit	**Fatal**	1	0	0	1	0	0	0	0	0	0	**2**
	Serious	4	2	3	2	3	4	6	7	4	9	**44**
	Slight	3	2	6	7	5	6	9	12	10	7	**67**
Following too close	**Fatal**	0	0	0	0	3	0	0	0	0	0	**3**
	Serious	2	5	4	6	3	5	4	6	7	3	**45**
	Slight	5	8	10	9	6	7	8	9	11	13	**86**
Defective Brakes	**Fatal**	0	0	1	0	0	0	1	0	0	0	**2**
	Serious	4	3	7	5	8	9	11	8	8	5	**68**
	Slight	6	9	12	12	13	9	10	8	6	7	**92**
Total		**25**	**29**	**43**	**42**	**41**	**40**	**49**	**50**	**46**	**44**	**409**

Risk 1: Accident Severity = Fatal; CF = Exceeding Speed limit. Marginal total = 2.

This is equivalent to scenario 2b with a CF with a High visibility / sensitivity rating and 2 factual variables where cells of size 1 and 2 could be unsafe.

If is required,

- ages 31-40 could be combined to give a cell total of 2 while not combining the other cells or
- ages 31-40 could be combined for all CFs and Accidents as in Table 7a

Table 7a: Age Group by Accident Severity for specified CFs

		Region A
First CF	Accident Severity	**Age 31-40**
Exceeding Speed Limit	Fatal	2
	Serious	44
	Slight	67
Following too Close	Fatal	3
	Serious	45
	Slight	86
Defective Brakes	Fatal	2
	Serious	68
	Slight	92
	Total	**409**

There are still low counts present. If combining ages is not sufficiently maybe the cell will need to be suppressed with another cell selected for secondary suppression.

Risk 2: Accident Severity = Fatal; CF = Following too close; All 35 year olds involved in the accident suffer a fatal injury. An intruder would know that a 35 year old friend or relation who was killed in this region died (according to a police officer) with a particular CF being influential. If this attribute is to be protected then

- Combine this CF with another CF or release the higher level non-specific CF
- Suppress the entire row
- Round the table

Risk 3: This is similar to Risk 1. Here there are low counts where Accident Severity = fatal and CF = defective brakes. As Defective brakes has not been classified as being high visibility/sensitivity the risk category is low indicating marginal totals need to be checked.
The risk now becomes subjective with knowledge of the data and uses the table maybe put to determining whether the table can be released without any protection.

More detail on lower geographies

Data can be release at lower geographical levels such as Local Authority as shown in the scenarios below. The populations at risk here is much smaller and there is a greater risk of disclosure. Information about accidents is in the public domain as this item from the BBC website on 6th December 2012 shows.

A cyclist has been killed in a collision with a lorry in east London.

The man was pronounced dead at the scene on Commercial Road, Stepney, shortly before 08:30 GMT.

His family have not yet been told of his death. The road is closed in both directions near the junction with Arbour Square.

Transport for London said 14 cyclists had been killed on the capital's roads this year. A police spokesman said investigations were under way.

An intruder will know the following:
Accident severity: Fatality
Type of vehicle: cyclist (presumably but not definitely pedal cyclist)
Road Type: Urban
Road Class: A road
Time: before 8:30am (but not too long before)
Local Authority: Tower Hamlets
Location: near junction of Commercial road and Arbour Sq

The report was updated later to state that the victim was a male in his 30s and a lorry driver had been arrested on suspicion of causing the death.

If detailed data are released for this LA especially ‘Time of accident’ and ‘Age of victim’ alongside CFs, it would not be too difficult for an intruder to link this published information above with a released table. Contributory Factors could be determined leading to a thorough account of the accident and allowing the user of the data to determine blame.
As an arrest has been made this might not matter in this case, but it could influence somebody (a juror?) if a case comes to court following publication of the data.

Does any disclosure risk identified represent a breach of public trust, law or policy: The National Statistics Code of Practice (Principle 5, Practice 1) is as below:

- ‘ Ensure that official statistics do not reveal the identity of an individual or organisation, or any private information relating to them, taking into account other relevant sources of information’

Any data which are released ought not to allow an individual to be identifiable i.e. they could be identified taking account of the ‘means likely reasonably to be used’. This should be considered when low counts are present in the data. It may be theoretically possible to identify an individual but in practice is would take an excessive amount of time and effort

Principle 5, Practice 4 states

- Ensure that arrangements for confidentiality protection are sufficient to protect the privacy of individual information, but not so restrictive as to limit unduly the practical utility of official statistics. Publish details of such arrangements.

A balance has to be struck between confidentiality and utility. There will be small element of risk in publishing any official statistics but a level of utility needs to be maintained to ensure these data are of use to researchers and the general public.

Select appropriate disclosure control methods to manage the risk: A summary of the methods recommended for each scenario are discussed in Table 6 in a previous section. Table redesign is the best approach to use, in particular recode the CF data to a higher level. Certain factual variables such as Single Age could also be recoded.
Recode on the basis of maintaining the greatest utility of the data. Any recoding ought to be carried out by people with knowledge of users' requirements. Some detail would be lost but this method avoids damaging the data, values published are genuine values.

Implement and disseminate: Always state that data are presented in this form due to the need for recoding caused by the application of statistical disclosure control.

5. Note on Outputs /Transparency agenda

This is a rich dataset and currently many tables are produced from it. Alongside the outputs published each September tables not currently published can be requested through the following approaches.

- General requests for bespoke tables
- FOI requests
- Parliamentary Questions

There is a possibility that if the number of publications increases, disclosure by differencing could be an issue especially if a large number of requests are made from the same source for tables with only slightly different characteristics.
Any new request ought to be compared against previous requests and releases to ensure a user is not attempting to obtain a number of linked tables which can be differenced in an attempt to identify individuals. Any requests for variables recoded into unusual categories should be looked at closely along with requests on slightly different locations (Police Authority and Local Authority maybe an example).

Releasing Microdata

Road traffic accident data can currently be downloaded from the UK data Archive through an End User Licence although no CF data is currently released in these data. Users are expected to adhere to certain conditions and one relating to confidentiality states:

To preserve at all times the confidentiality of information pertaining to individuals and/or households in the data collections where the information is not in the public domain. Not to use the data to attempt to obtain or derive information relating specifically to an identifiable individual or household, nor to claim to have obtained or derived such information. In addition, to preserve the confidentiality of information about, or supplied by, organisations recorded in the data collections. This includes the use or attempt to use the data collections to compromise or otherwise infringe the confidentiality of individuals, households or organisations.

If more detailed microdata are to be released including CF data these are likely to be personal data allowing individuals to be identified. A more stringent licence could be developed by DfT to define restrictions on publications and any penalties if confidentiality guidelines are broken.

This licence could be equivalent to the Special Licence which allows specific ONS microdata sets to be released. There would be different conditions for a licence developed by DfT as only the ONS is subject to the Statistics and Service Registration Act (2007) Section 39 which allows personal data to be made available to Approved Researchers.

6. State the guidelines

Table 6 gives a detailed summary of the disclosure control methodology to protect tables produced from the DFt road accident dataset. The main steps to follow are:

- What is the risk category of the data? In particular
  - level of geography
  - number of factual variables
  - visibility / sensitivity of the Contributory Factor
- What action needs to be taken
  - protect unsafe cells
  - protect unsafe marginal totals
- What disclosure control methodology (if any) needs to be applied
  - table redesign, e.g. combine categories
  - table rounding, to a specified base (e.g. 3 or 5) so that all cells are multiples of these values
  - suppression of cells, both primary and secondary

7. Updating the advice

Variables can change but only slightly and in line with the quinquennial review. Therefore it is unlikely that the advice given in this document will require significant changing over time.
This document could be checked during the review to see if alterations are required due to changes to variables or rewrites of relevant laws or Code of Practice/National Statistician guidance.

References

Freedom of Information Act (2000)

http://www.legislation.gov.uk/ukpga/2000/36/contents

Freedom of Information (Scotland) Act 2002

http://www.legislation.gov.uk/asp/2002/13/contents

UK Statistics Authority Code of Practice for Official Statistics

http://www.statisticsauthority.gov.uk/assessment/code-of-practice/index.html

National Statistician's guidance: Confidentiality of Official Statistics

http://www.statisticsauthority.gov.uk/national-statistician/guidance/index.html

GSS Standards for Administrative data (to be revised later in 2013)

http://www.ons.gov.uk/ons/guide-method/best-practice/disclosure-control-policy-for-tables/index.html

Appendix 1

Factual variables

When an accident is recorded a large amount of information defined as factual is collected. This can refer to the accident, (road conditions, light conditions), the vehicle (type of vehicle, junction location of vehicle) and casualty (age, pedestrian location).

The list in this appendix includes variables which are included in current tables and those which could be of most interest if included in future releases. It is not an exclusive list and other factual variables could be included in outputs.

Variable	Category	Recodeable. Possible options shown
Region / LA		
Road user type	Pedestrian Pedal Cyclist Motorcycle Rider/Passenger Car Occupant Bus or Coach Occupant Van/Light Goods Vehicle Occupant HGV Occupant Other Vehicle Occupant All road Users of which children	Pedestrian, Pedal cyclist, Motor cyclist, Car, Other vehicle
Age of road user type	Single Age	5yr bands or 10 yr bands
Sex	Male Female	
Road type (Closely related to Road Class)	Urban Rural Motorway	
Time	Hour	
Severity of Accident	Fatal Serious Slight	Fatal Serious/slight
Road Class	Motorways A road B road Other	
Vehicle Type	Pedal cycle Motor cycle Car Bus or Coach LGV HGV	
Speed Limit (actual or banded as shown) Can Built up / no built up be based on these speed limits?	1 – 20 mph 21 - 30 mph 31 – 40 mph 41 – 50 mph 51 – 60 mph 61 – 70 mph Missing, 0 or over 70 mph	These can be banded as required
Urban/ Rural indicator (closely related to Road Type and Road Class)	Urban Rural Unallocated Undefined	
Breath Test	Positive	

	Negative Not Requested Refused to provide Driver not contacted Not provided	

Appendix 2

Contributory Factors

High level	Low level	Visibility (how obvious would this be to a passerby / neighbour etc)	Sensitivity (i.e. how much damage will this cause to the motorist if released
Road Environment Contributed	Slippery road (weather)	High	Medium
	Deposit on road (e.g. oil, mud, chippings)	High	Medium
	Poor or defective road surface	High	Medium
	Sunken, raised or slippery inspection cover	High/medium	Medium
	Road layout (e.g. bend, hill, narrow carriageway)	High	Medium
	Temporary road layout (e.g. contraflow)	High	Medium
	Animal or object in carriageway	High/Medium	Medium
	Inadequate or masked signs or road markings	High/Medium	Medium
	Defective traffic signals	High	Medium
	Traffic calming (e.g. speed cushions, road humps, chicanes	High	Medium
Vehicle Defects	Tyres illegal, defective or under-inflated	Low	High
	Defective lights or indicators	Medium	High
	Defective brakes	Low	High
	Defective steering or suspension	Low	High
	Defective or missing mirrors	Medium	High
	Overloaded or poorly loaded vehicle or trailer	High	High
Injudicious Action	Following too close	High	High
	Exceeding speed limit	High	High

	Disobeyed Give Way or Stop sign or markings	High	High
	Disobeyed automatic traffic signal	High	High
	Travelling too fast for conditions	High	High
	Cyclist entering road from pavement	High	High/ Medium for motorist
	Illegal turn or direction of travel	High	High
	Disobeyed pedestrian crossing facility	High	High / Medium for motorist
	Vehicle travelling along pavement	High	High
	Disobeyed double white lines	High	High
Driver/ Rider Error or Reaction	Failed to look properly	Medium	High
	Failed to judge other person's path or speed	Medium	High
	Poor turn or manoeuvre	High	High
	Sudden braking	High	High
	Swerved	High	High
	Junction overshoot	High	High
	Junction restart (moving off at junction)	High	High
	Failed to signal or misleading signal	High	High
	Too close to cyclist, horse or pedestrian	High	High
	Loss of control	High	High
Impairment or Distraction	Impaired by alcohol	Low / Medium	High
	Impaired by drugs (illicit or medicinal)	Low / Medium	High
	Driver using mobile phone	High	High
	Fatigue	Low / Medium	High
	Distraction in vehicle	Low / Medium	High
	Distraction outside vehicle	Medium	High
	Illness or disability, mental or physical	Low	High

	Uncorrected, defective eyesight	Low	High
	Rider wearing dark clothing	Medium	Medium
	Not displaying lights at night or in poor visibility	High	Medium
Behaviour or Inexperience	Careless, reckless or in a hurry	Medium	High
	Learner or inexperienced driver/rider	Low	High
	Aggressive driving	High	High
	Nervous, uncertain or panic	Medium	High
	Unfamiliar with model of vehicle	Low	High
	Inexperience of driving on the left	Low	High
	Driving too slow for conditions or slow vehicle (e.g. tractor)	High	High
Vision Affected by	Stationary or parked vehicle(s)	High	High
	Road layout (e.g. bend, winding road, hill crest)	High	Medium / Low
	Dazzling sun	High	Medium
	Rain, sleet, snow or fog	High	Medium
	Spray from other vehicles	High	Medium
	Dazzling headlights	Medium	Medium
	Vehicle blind spot	Medium	Medium / Low
	Vegetation	High	Medium / Low
	Buildings, road signs, street furniture	High	Medium / Low
	Visor or windscreen dirty, scratched or frosted etc.	Low	High
Pedestrian Only (Casualty or Uninjured)	Failed to look properly	High	High
	Careless, reckless or in a hurry	High	High
	Failed to judge vehicle's path or speed	Medium	High
	Crossing road masked by stationary or parked vehicle	High	High
	Impaired by alcohol	Low / Medium	High
	Impaired by drugs (illicit or medicinal)	Low / Medium	High

	Dangerous action in carriageway (e.g. playing)	High	High
	Wrong use of pedestrian crossing facility	High	High
	Pedestrian wearing dark clothing at night	High	High
	Disability or illness, mental or physical	Low	Medium
Special Codes	Stolen vehicle	Low	High
	Vehicle in course of crime	Low	High
	Emergency vehicle on a call	High	Medium / Low
	Vehicle door opened or closed negligently	High	High
	Other – Please specify below	High/medium/Low	High/medium/Low