Department for Transport

Linking Police and Hospital data on Road Accidents in England: 1999 to 2009 results

February 2012

Contents

Summary

The main source of road casualty statistics in Great Britain is the STATS19 database of accidents reported to police. More recently, it has become possible to identify road traffic casualties admitted to hospital in England from the Hospital Episode Statistics (HES) database. This provides an alternative source of information on more seriously injured casualties.

The linkage can provide insight into the completeness of police data, and brings together more detailed information on accident circumstances and medical outcomes for further research. This report summarises the methodology used to link records from the two datasets, and some subsequent analysis of the linked data.

Feasibility of linkage

STATS19 and HES lack a unique common identifier, so a rules-based method was developed to link the two datasets based on partial identifiers including age, sex, date and region of accident and admission, casualty postcode and road user type.

As the values of linking variables are missing or incorrectly coded in some cases, perfect identification of correct matched and non-matched record pairs is not possible. Initial attempts to assess the quality of linkage suggest a small proportion (around 3%) of incorrectly linked records, but that up to 20% of genuine matches may have been missed. The overall quality of linkage improves over time, in particular as the recording of casualty postcode on STATS19 becomes more complete.

A linked dataset of around 190 thousand records covering years 1999 to 2009 was created. Overall, around a third (32%) of HES records were linked to STATS19, and 37% of STATS19 serious records were linked to HES. However, due to the likelihood of missed matches these proportions are likely to under-state the true overlap between the datasets.

The large sample size of the matched dataset should make it suitable for a range of analyses. The possibility of errors in linkage and in classification of road casualties in the police and hospital datasets should be kept in mind, though these are unlikely to invalidate the broad findings.

Factors associated with levels of reporting to police

Comparing linked and unlinked hospital records provides information on factors associated with the likelihood of road casualties coming to the attention of police.

- Overall, 41% of traffic accidents admissions within the scope of STATS19 recorded in HES are linked to STATS19. This rises to 48% for traffic accidents excluding non-collision pedal cycle accidents. Given the likely underestimation of the number of records linked, this suggests that over half of those admitted are recorded in STATS19

- Excluding non collision pedal cycle accidents, the proportion of admitted casualties linked to STATS19 is similar across the main road user groups, although bus occupant casualties may be under-represented. A very low proportion of pedal cyclists injured in non-collision accidents become known to police, but in collision accidents the proportion of pedal cyclist casualties linked is comparable with other vulnerable road user groups

- STATS19 appears broadly representative in terms of age and gender of casualties, although child pedestrian casualties may be over-represented relative to other groups

- Among casualties admitted to hospital those with more severe injuries are more likely to be known to police.

Factors associated with hospital admission

Comparing linked and unlinked STATS19 records provides information on the types of casualties known to police that are more likely to be admitted to hospital. Results suggest pedestrians and motorcycle users are more likely to be admitted relative to other road user groups, and children and elderly casualties relative to other ages.

In broad terms, the proportion of STATS19 records linked to HES is correlated with the proportion of non-fatal casualties coded seriously rather than slightly injured in STATS19.

Classification of injury severity

Analysis of linked records allows the accuracy of coding of injury severity by police to be explored. In particular, the definition of serious injury used by police means that all those casualties appearing in HES should be coded seriously injured in STATS19.

- Overall, 58% of linked casualties are correctly coded as serious with the remainder being coded slight. This proportion is considerably higher (around 80%) among casualties spending longer in hospital or with more severe injuries. It is possible that some of those misclassified by police as slightly injured have relatively minor injuries but are admitted to hospital for observation.
- There is no clear evidence of a systematic deterioration in the coding of injury severity over the past decade, based on these results. Although the proportion of linked records coded serious falls marginally, reflecting a fall in serious injuries among non-fatal casualties in the whole STATS19 dataset.
- The misclassification of injury severity by police is higher for those casualties where the primary diagnosis in HES is classed as a ‘superficial’ injury (32% coded serious) or a dislocation, sprain or strain (42%) particularly when to the neck. Conversely, 74% of those with fractures and internal injuries are correctly classified as serious.

It is also possible to derive alternative measures of injury severity from the HES data, for example based on length of stay in hospital or MAIS level. These suggest that motorcyclists have more severe injuries and spend longer in hospital, on average, than other road user groups.

Estimating total serious casualties using capture-recapture

Although the data linkage provides some information on the completeness of the police data, neither STATS19 nor HES provides full coverage of serious road casualties. To assess the proportion of the total contained in each source, a method known as capture-recapture was applied. This method relies on a number of assumptions, some of which are unlikely to hold in this context, so conclusions should be considered broadly illustrative rather than precise estimates. Bearing the limitations in mind:

- Overall, around a third of estimated total serious road casualties (according to the definition used by police) are likely to become known to police and included in STATS19 as serious casualties with around 40% admitted to hospital and included in HES as road traffic accidents

- There was no strong evidence that the estimated proportion of total serious casualties known to police has changed over the last decade. The proportion admitted appears to have steadily increased, perhaps as a result of changing hospital practices.

- As a broad illustrative estimate suggests that the total number of road casualties (including those not reported to the police or admitted to hospital) in Great Britain, excluding pedal cycle casualties in accidents with no motor vehicle, is of the order of 85 thousand in 2009. This is in line with equivalent estimates published by DfT.

- The trend over time in estimated total casualties is more similar to that shown by STATS19 than HES, supporting the conclusion that the police data, although incomplete, is a reliable measure of trends at the national level.

Future work

This work presents an initial analysis of the linked data. However, the value of linking police and hospital data will only become evident when the data have been used more widely to explore the medical consequences of road accidents in conjunction with the information on accident circumstances.

This analysis has looked at individual variables in turn, and a more sophisticated approach might be to consider a multivariable analysis. There are some areas which might be explored further, such as considering regional variations, though this is unlikely to be straightforward.

Although a sufficiently robust linkage has been developed, there is scope for some improvement and for extension, possibly to other countries or, if possible, to Accident and Emergency data which is currently available but of insufficient quality.

Conclusions

This study illustrates that linkage of police and hospital inpatient data for England over a number of years is feasible, although imperfect. The findings are useful in illustrating the strengths and limitations of the two sources.

The majority of the broad results of this work are not new, so their main value is in supporting existing conclusions based on more localised or older studies.

Given the limitations of the data, it is difficult to draw firm conclusions, and alternative interpretations are possible. However, these results add to the evidence base regarding the coverage and representativeness of STATS19 data in particular, suggesting that it remains a suitable source for monitoring patterns and trends at national level.

1. Introduction

Information on casualties in road traffic accidents in England is available from the long established data collected by the police (known as STATS19[1]), and more recently from data on hospital admissions (HES - Hospital Episode Statistics, collated by the NHS Information Centre and supplied for all hospitals in England). These two sources provide alternative, though not equivalent, measures of the number and trend of seriously injured casualties on the roads of England.

In recent years, following a review of road casualty statistics by the UK Statistics Authority (UKSA) [26] attempts have been made to estimate the total number of road casualties taking account of information from a range of sources. This includes survey data and hospital records, and analysis of hospital admissions for road casualties alongside statistics derived from police data [7, 24].

While STATS19 provides the most useful single source of data on road accidents and form the basis for published statistics [e.g. 24], it has long been known that they provide an incomplete record and that a considerable proportion of non-fatal road accidents do not become known to police.

The STATS19 and HES data have shown differing trends particularly during between 2002 and 2005 [e.g. 1, 2 , 3 ,21,24], with police data showing falls in seriously injured casualties compared to increased road casualties admissions to hospitals. This difference may occur for a combination of reasons including definitional differences, changes in levels of reporting to the police or police recording practice, and changes in hospital admission practices. Although these are not fully understood, a recent report by the UKSA accepted that the fall in serious casualties is likely to be genuine [25].

Recently, the Department for Transport (DfT) in collaboration with the NHS Information Centre have linked STATS19 records of people injured in road accidents in England with HES records of patients admitted to hospital who were injured in a road accident at individual record level. This work has two aims:

- To provide, as far as possible, a deeper understanding of the completeness and trends shown by the two datasets
- To create a linked dataset as a resource for further research into road accidents, bringing together the information on accident circumstances (in STATS19) with medical consequences (from HES)

This report focuses on the first of these aims, and provides details of the linkage methodology. A brief introduction to the two data sources and the variables available for linkage are described in section 2. Section 3 contains details of the methodology, including a section on the quality of the matching process. Section 4 presents comparisons of the resulting linked data with both HES and STATS19 to explore factors affecting levels of reporting to police and propensity to attend hospital following an accident. Section 5 considers how far the linked data can be used estimate the number of ‘seriously injured’ casualties and look at trends over time, given the inherent limitations.

[1] Named after the number of the first questionnaire issued when the system was introduced in 1949

With regards to the second aim, the Department will make the raw linked data available to researchers. Please contact roadacc.stats@dft.gsi.gov.uk with a request for information, providing evidence of previous research projects/publications, information about the intended use of the data and the purpose of the research. Researchers are required to have an End-User Licence for the STATS19 data and a HES Data Re-Use Agreement to access to the data.

While there has been previous linkage of STATS19 data with local hospital records for particular geographical areas [e.g. 3, 4, 19], and for other countries [e.g. 5, 10 for Scotland], this work represents the first attempt at such linkage covering the whole of England. Compared with previous British studies, it typically provides greater coverage in terms of both geography and time period. However as the HES inpatient dataset covers only those casualties admitted to hospital (and not those, for example, attending A&E only), this linkage tells us relatively little about those less severely injured. Therefore it adds to the existing evidence, rather than providing the full picture.

2. Data Sources

2.1 The STATS19 file

All personal injury road accidents on public highways involving at least one vehicle, reported to and recorded by the police (within 30 days of occurrence) appear in the Department for Transport (DfT) national road accident database, known as STATS19.

The scope and detail of STATS19 allows the identification of different accident circumstances, enabling road safety policies to target appropriate interventions to reduce the number of accidents and their resulting casualties. Some 50 data items are collected for each accident, recording information on the accident, the vehicles involved and the casualties[2].

Casualties are classified as fatal (death within 30 days of the incident), seriously injured or slightly injured. In STATS19, the definition of serious injury includes all casualties admitted to hospital and certain injuries, such as fractures, regardless of whether or not the casualty was admitted to hospital (see glossary for definition). The severity of casualty is recorded by the reporting police officer, usually on the basis of information available within a short time of the accident. Guidance to reporting police officers is contained in the document STATS20 [13]

On average, the STATS19 data analysed consists of around a quarter of a million records each year between 1999 and 2009. The reported number of non-fatal casualties resulting from road accidents in England is shown in Table 2.1.

Fatal casualties were not included in the data for linking since the majority (around 80%) of fatalities in road accidents die before admission to hospital [3]. Only data from 1999 onwards were included in the matching process because this is the first year in which casualty postcode information was recorded in STATS19.

Table 2.1: STATS19 records used in linking by year and severity: England 1999-2009

	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Serious	33,710	32,951	32,176	31,285	29,292	27,057	25,210	24,856	24,218	22,246	21,326
Slight	248,494	249,855	244,586	234,755	225,603	218,991	212,539	201,026	192,744	179,788	173,574
All	**282,204**	**282,806**	**276,762**	**266,040**	**254,895**	**246,048**	**237,749**	**225,882**	**216,962**	**202,034**	**194,900**

The variables used for linking from the STATS19 file included:

- Age and gender of casualty
- Road user type (car, motorcycle, pedestrian, pedal cyclist or other)
- Casualty class (driver/rider, passenger or pedestrian)
- Casualty home postcode
- Local authority district of accident
- Date of accident

[2] A copy of the STATS19 form can be found at: http://assets.dft.gov.uk/statistics/series/road-accidents-and-safety/stats19-road-accident-injury-statistics-report-form.pdf

- Strategic health authority region of accident (derived from the accident location)

The road user type and casualty class variables were combined to create a combined ‘road user class’ variable for use for the linking process as described in section 3. The categories for road user class are shown in Table 2.1[3].

Table 2.1: STATS19 class of road user

Code	Road user class
1	Driver of motor vehicle
2	Passenger of motor vehicle
3	Rider of motorcycle
4	Passenger of motorcycle
5	Pedal cyclist
6	Pedestrian
7	Other or unknown

2.2 The HES file

Information on casualties admitted to hospital as in-patients in England is contained in the Hospital Episodes Statistics (HES) database owned by the Information Centre of the National Health Service (NHS). It is compiled by the Information Centre from over 300 NHS Trusts in England, and is an administrative return used in reimbursing hospitals for work completed against contracts.

Casualties treated in Accident and Emergency departments who are not subsequently admitted to a hospital are not included in the HES database. However, all casualties admitted to a bed in a hospital in England should be recorded in the data even if the admission did not require an overnight stay.

The main unit of recording in the HES data represents an episode of care under a particular consultant and contains clinical details of the patient’s condition coded to the International Classification of Diseases, 10th edition (ICD-10)[4]. The ICD-10 is an international standard diagnostic classification used in health records. Under each episode, a patient can have up to 20 diagnoses relating to their condition (although this has varied over time[5]). Other recorded information includes admission to and discharge from hospital. Further details about the HES data can be found on the website www.hesonline.nhs.uk/ (see also [22, 23] for general guides to the HES data).

The ICD-10 codes of particular interest for this project included patients with a diagnosis code of an *external cause of injury – subgroup of transport accidents* (ICD-10 V codes, Chapter XX). The transport accident codes (V01 to V89, excluding V81) allow the identification of road transport accident casualties. More specifically, they allow the identification of road user type and casualty class (e.g. casualty being a passenger of a motorcycle). In addition, they allow the identification of road casualties of traffic accidents (vehicle accidents occurring on the public highway).

[3] This is based on a similar approach used in the matching of police and hospital data for Scotland [10]

[4] ICD-10 Reference http://www.who.int/classifications/apps/icd/icd10online/

[5] HES recorded up to 7 diagnoses to 2001/02, up to 14 to 2006/07 and up to 20 from 2007/08 onwards.

Further, the locations and types of the subsequent injuries can be identified for any patients who were diagnosed with ICD-10 codes relating to *injury, poisoning and certain other consequences of external causes* (ICD-10 S and T codes, Chapter XIX)

The extract used for linking is selected on the *external cause of injury* for all HES records. The criteria are to select those patients who have an external cause of injury relating to a road transport accident (codes V01 to V89, excluding V81). Non-traffic accidents, e.g. any vehicle accident that occurred entirely in any place other than a public highway (please see glossary for further details), as recorded in the V codes are outside of the scope of STATS19. However, the extract did not distinguish between traffic and non-traffic accidents to allow for miscoding of external causes of injury by hospitals.

As with the STATS19 extract, any fatal casualties (those recorded as having died in hospital) were excluded. Please note that hospital deaths, including casualties who died after 30 days of the road accidents were excluded, even though these deaths would be classified as seriously injured in STATS19. However, the proportion of all deaths which occurred after 30 days was small (around 5% of all emergency hospital admission deaths that were within the scope of STATS19 road accidents).

The HES extract also excludes elective (i.e. planned, non emergency) admission to exclude repeated admissions to hospital after a road accident.

2.2.1 Removal of records relating to same patient and accident

Hospital admission records are based on periods of care (‘episodes’) under a particular consultant, so patients can be counted more than once (e.g. if they transfer to another consultant). Episodes join together to form ‘spells’, with each spell representing care under one hospital provider.

A single patient may therefore have more than one episode (or spell) of care arising from a single accident. Therefore some data cleaning (de-duplication) was required to identify records relating to the same patient and the same accidents. However, there is also the possibility that a patient may have multiple admissions as a result of involvement in more than one accident. For matching purposes the former cases (multiple emergency admissions from a single accident) should not be reduced to a single record while the latter cases should be consolidated where possible.

The recording of incomplete or contradictory data further complicates this process as all data relevant to a spell will not necessarily be recorded in each of its episodes. Diagnostic and cause codes in particular are not necessarily recorded in every episode referring to the same hospital case or spell.

This de-duplication was carried out prior to matching to the STATS19 data. First, all episodes for the same patient were grouped together by chronological order of episode start date. If there were more than 14 days between the end of one episode and the start of the next, then this was assumed to be related to two separate accidents (this is likely to be conservative but has only a marginal effect on the number of records)[6]. For the linkage process, only the episode with the earliest date was selected for each accident.

Table 2.3 shows the number of records in the hospital data file before and after this de-duplication process. In following sections, the number of records not flagged as duplicates will be taken as the number of HES records.

[6] This approach is imperfect but reasonable given the data provided. Better algorithms for consolidation of episodes to spells exist and we will explore the potential to apply these in any future linkage

Table 2.3: Number of records in HES extract before and after removing duplicates[7]

	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Total HES records	57,350	56,108	55,005	53,603	56,459	57,431	61,239	60,315	59,890	59,469	63,115
Not flagged as duplicate	52,347	51,093	50,078	49,580	53,023	54,339	57,876	56,826	56,537	56,246	59,955
Flagged as duplicate	5,003	5,015	4,927	4,023	3,436	3,092	3,363	3,489	3,353	3,223	3,160
Proportion used in linkage	*91*	*91*	*91*	*92*	*94*	*95*	*95*	*94*	*94*	*95*	*95*

2.2.2 Description of variables

The variables used for linking from the HES file were:

- Age and gender of patient
- Patient home postcode
- Strategic health authority of hospital (using the 28 areas existing pre-2006)
- Local authority district of hospital
- Local authority district of the patient
- Date of admission

In addition, details of the road user type and casualty class relating to the patient can be derived from the *external cause of injury – transport accidents* coding as already discussed (ICD-10 V codes, Chapter XX). The casualty class and road user type relating to each V code was extracted and used to create a combined ‘road user class’ variable with the same categories as the ‘road user class’ variable constructed for STATS 19 casualty as shown in Table 2.1.

Other variables of interest from the HES database include the diagnoses of injury or illness (derived from the ICD-10 S and T codes*, injury, poisoning and certain other consequences of external causes*, as discussed above), and the length of stayed in hospital[8]. Table 2.4 provides information about the recording of external causes of injury (V codes) and the injury diagnoses (S and T codes). On average each record has 1.6 injury-related diagnoses. The types of injuries sustained are explored further for the matched dataset, and the results can be found in section 4 of this report.

Table 2.4: HES diagnosis field completeness (all records including duplications)[5]

	Injury, poisoning and certain other consequences of external causes	External cause of injury – transport accidents	Other non-injury diagnoses	Blank	Total

[7] In addition, a small number of records with admission date 1st or 2nd January 2010 were included in the file for linking, to allow for matches with accidents on 31st December 2009 (see description of methodology below)

[8] In the data file provided, this information was only available from cases where the spell in hospital consisted of a single episode of care. In cases where the patient had a further episode of care, the length of spell data is missing.

	S code	*T code*	*V code*			
Diagnosis 1	582,698	11,918	0	45,565	0	**640,181**
Diagnosis 2	188,122	7,536	416,907	27,616	0	**640,181**
Diagnosis 3	132,310	7,959	121,734	122,484	255,694	**640,181**
Diagnosis 4	62,480	4,604	57,069	98,023	418,005	**640,181**
Diagnosis 5	31,804	2,931	25,224	68,291	511,931	**640,181**
Diagnosis 6	15,674	1,821	12,963	44,753	564,970	**640,181**
Diagnosis 7-20	15,333	2,751	11,287	66,991	8,866,172	**8,962,534**
All diagnoses	**1,028,421**	**39,520**	**645,184**	**473,723**	**10,616,772**	**12,803,620**
Average per record	*1.6*	*0.1*	*1.0*	*0.7*		

Table 2.5 shows the distribution of number of injury diagnoses recorded. Around 5% of records have no injury diagnosis. This could be because they were genuinely uninjured[9] or the result of inaccuracies in coding. Over three quarters of HES records have either one or two injury diagnoses, and less than 1% have more than 6 diagnoses.

Table 2.5: Distribution of number of injury diagnoses for HES records (all records 1999-2009, including duplications)

Number of injury diagnoses (S or T codes)	**Total**	***Percent***
0	33,058	5.2
1	362,034	56.6
2	132,969	20.8
3	58,671	9.2
4	27,228	4.3
5	13,453	2.1
6	7,547	1.2
More than 6	5,221	0.8
Total records	**640,181**	**100**

[9] Previous studies suggest that a small proportion of casualties recorded by police have no injuries (see for example [12])

3. Linkage method and results

3.1 Background

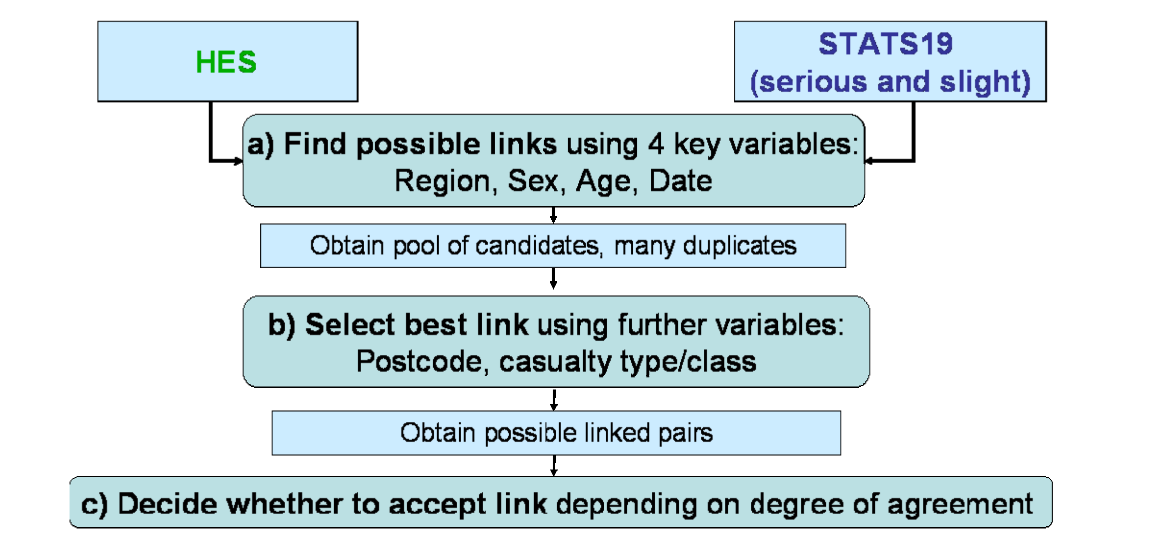
The police and hospital datasets share a large number of identical variables and some common partial identifiers but lack a common uniquely identifying variable. This forces the use of the common characteristic variables, in particular sex, age, date (of admission to hospital and of accident) and region (of accident and of hospital) to find possible matched pairs. Further variables, most notably postcode, can then be used to identify the most likely correct match from the candidates generated, according to a set of rules. The approach is therefore a mix of exact and rules-based linkage rather than formally probabilistic (in particular agreement weights are not applied).

The first linkage of STATS19 and HES was developed by the Office for National Statistics, with initial results published in 2008 [7]. Subsequently, the results were peer reviewed and it was apparent that the number of false positive links (i.e. incorrect linkages made for two non-matching records) was unacceptably high. As a result, changes were made to the linkage rules and the revised method and results are presented in this section.

3.2 Method

Linking was performed using (Proc) SQL commands within the SAS programming language, on a laptop with 2.4GHz Intel Core 2 Duo processor and 3.5Gb RAM running Windows XP.

Figure 3.1 Overview of linking process


HES
STATS19 (serious and slight)
a) Find possible links using 4 key variables: Region, Sex, Age, Date
Obtain pool of candidates, many duplicates
b) Select best link using further variables: Postcode, casualty type/class
Obtain possible linked pairs
c) Decide whether to accept link depending on degree of agreement

In summary, the linkage process (shown diagrammatically in Figure 3.1) was as follows:

1 Generate candidate linked pairs on the four key variables: age, sex, date and Strategic Health Authority (SHA). This will result in a many-to-many relationship containing many duplicates (i.e. each HES record will be linked to many

STATS19 records, and some STATS19 records will be linked to more than one HES record).

2. Resolve duplicates using the remaining variables (postcode, casualty type/class). This was done by applying a set of rules to select the pairs most likely to represent a true match from among the pool of potential candidates generated.
3. For the resulting linked record pairs, determine whether or not they represent a true match with sufficient confidence, based on the degree of agreement on the linkage variables.

3.2.1 Preparation of data files for linkage

Before linking was carried out, variables on each file were standardised to allow comparison. For example, children aged 0 years are coded differently in the hospital data and therefore must be recoded to be consistent with the STATS19 dataset. As discussed previous, a road user class variable was also derived in both datasets.

3.2.2 Generation of candidate links

The first stage of the linkage involves generating candidate links, using four key variables – age, gender, date and region (SHA). The general aim is to produce a manageable number of possible linked pairs, without excluding too many correct matches. If the conditions are defined too tightly it is likely that too many correct matches would be missed. Conversely, if the criteria are too weak, there is more chance that incorrect linkages are generated (and therefore more computational work is needed to identify which of the candidates is the most likely true match).

Tolerance is allowed in the level of agreement on these variables (except for gender), to allow for recording inaccuracies or genuine differences between the two data sources. These tolerances were developed by ONS using a trial and error approach. Essentially this acts as a blocking process for the matching process, breaking the files into subsets. Records for the same person are unlikely to appear in different blocks. Only records within the same blocks are compared, thereby reducing the number of comparisons and resulting computation effort. The tolerances used for included variables are described below.

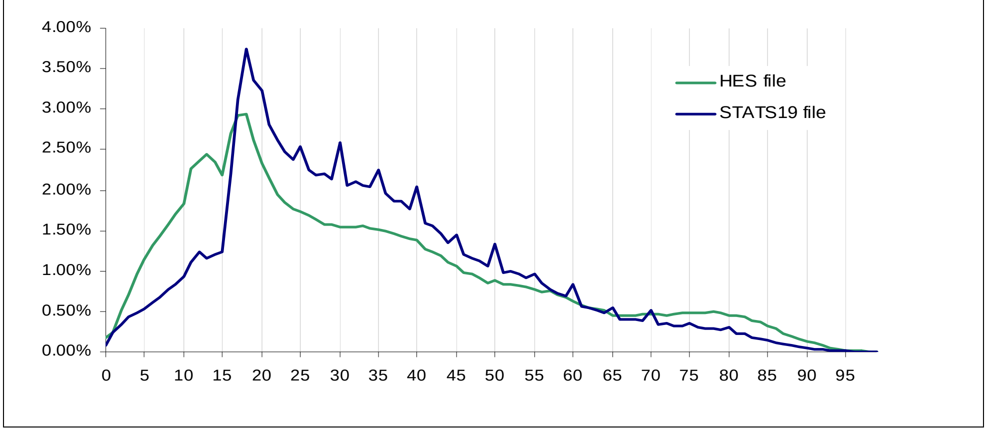
Gender: exact agreement required, as this is considered unlikely to be miscoded.

Date: date of admission (HES) were allowed to be up to two days after date of accident (STATS19) as it is possible that a casualty may not be admitted immediately following an accident. For example, in cases where an accident happens during the evening or where symptoms are not immediately apparent.

Age: the distribution in the STATS19 file suggests that age is sometimes estimated by the police officer at the scene of an accident (see Figure 3.2, with noticeable heaping at ages ending in 0, and to a lesser extent 5). In contrast, the age variable in HES is derived from the date of birth held on patient records and is therefore likely to be more accurate. The following tolerances were allowed in linking on ages over 20:

- where the STATS19 age ends in 0 or 5, up to 3 years difference either way is allowed
- in other cases, 1 years difference is allowed.

Figure 3.2 STATS19 and HES age distributions (1999-2007 data)



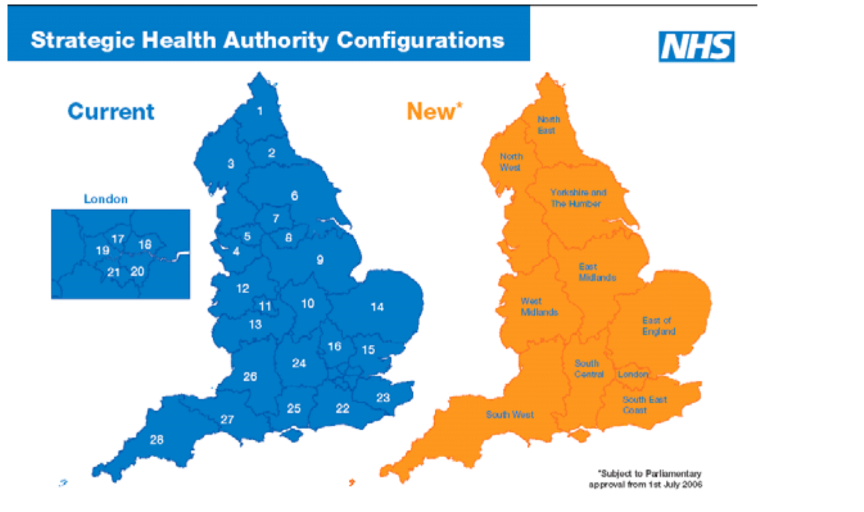
Region (SHA): matching between neighbouring SHAs were permitted to allow the possibility that casualties of road accidents may be taken to a neighbouring SHA for treatment[10].

Over the linkage period, the SHA boundaries have changed with successive re-organisations of the NHS. A lookup was created linking each lower-level health authority of treatment to the 28 SHAs extant up to June 2006 and also to the 10 SHAs created in a re-organisation in July 2006. Exploratory work by ONS found that the 28 areas serve as conveniently sized geographic units for blocking purposes whilst the 10 areas were too broad. Figure 3.3 shows the SHA regions used for the linking process.

The result of the first stage of the linkage was that, on average, around 10 possible links in the STATS19 file were found for each HES record, within the allowed tolerances.

[10] Note that cases where a casualty is admitted to a non-neighbouring SHA to that where the accident took place will not be linked in this process. Although this inevitably means some true matches will be missed, exploratory work shows these to be small in number (of the order of under 100 per year) and excluding them reduces computational effort.

Figure 3.3 Strategic Health Authority regions: 2005 configurations as used in linkage



3.2.3 Identification of most likely matched pairs

Having identified many possible candidate linkages for each record in the hospital file, the next stage of the process was to choose the pair most likely to represent a true match from amongst them. This is done through applying a set of rules, based on the exactness of the initial linkage and agreement on further variables including postcode, local authority area of the road accident compared to the residing local authority of the patient (LAD) and road user class.

Postcode is a particularly useful variable for this de-duplication as it has high power to discriminate between records (on average there are around 15 addresses per postcode, which is powerful here given the relatively small proportion of the population that are road casualties in any given year). Taken together with age, sex and date it is even more useful[11]. However the recording of postcode on the STATS19 file is incomplete, increasing from around 40% of records in 1999 to over 80% after 2006 (85% in 2009), so these variables cannot be used to identify all true matches and further rules are needed.

Individuals were assumed have a higher likelihood of being involved in a road accident near their homes compared to other places. So more weighting was given to matches where the residing local authority of the patient (from HES records) and the local authority where the road accident occurred (from STATS19 records) were the same.

In total, 24 levels of agreement are defined, as shown in Table 3.1. These levels are essentially subjective and defined after some trial and error. They are ordered with level 1 representing the strongest linkage (that is, those most likely to represent correctly matched records belonging to the same person), down to level 24 where there is a greater degree of disagreement. Any cases where the degree of agreement between the linked STATS19 and HES records does not fall into one of these categories are unlinked, coded 99 are considered to be non-matches.

[11] For example, in the 2007 STATS19 data, of 176,624 seriously or slightly injured road accident casualties in England with a valid postcode recorded, there were 175,082 distinct combinations of postcode, sex, age and date present

Once each candidate link is assigned with a level, for each HES record the STATS19 record with the lowest level link is selected as the most likely to represent a true match.

There remains the possibility that the same STATS19 record may be identified as the best candidate link for more that one HES record (i.e. the initial stage of matching results in a many to many relationship). Thus, a further de-duplication process is required, essentially carried out in the same way but ordering the HES records linked to the same STATS19 record by agreement level. The result is a set of one to one links between the two datasets with a level assigned according to degree of agreement on linkage variables.

Note that in some cases it was not possible to distinguish between two (or more) candidate links – they may have the same level of agreement (particularly where this is relatively weak). Such cases were flagged as unresolved and treated as unlinked. In practice one of the cases may represent a true match, so this probably means that the number of linkages achieved is a slight underestimate.

3.2.4 Determination of number of links considered to be true matches

The final stage of the process is to determine which of the linkages represent correctly matched records (that is, belonging to the same person). Any linked pair where the agreement level is below 99 is taken to be matched. In practice this will result in both incorrectly linked records, and missed matches – the likely extent of these are briefly discussed below.

Table 3.1 Summary of agreement levels

Level	SHA	Age	Date	Postcode	RU class*	LAD
Valid STATS19 postcode						
1	Exact	Exact	Exact or HES + 1	Exact - all chars	Exact	Exact
2	Exact	Exact	Exact or HES + 1	Exact - all chars	NA	NA
3	Adjacent	Exact	Exact or HES + 1	Exact - all chars	Exact	Exact
4	Adjacent	Exact	Exact or HES + 1	Exact - all chars	NA	NA
5	Exact or adjacent	Exact or Fuzzy	Exact or upto HES + 2	Exact - all chars	Exact	Exact
6	Exact or adjacent	Exact or Fuzzy	Exact or upto HES + 2	Exact - all chars	NA	NA
STATS19 postcode present (but may not be valid)						
7	Exact	Exact	Exact or HES + 1	First part + 2/3 of last part OR First 2 chars + last part	Exact	Exact
8	Exact	Exact	Exact or HES + 1	First part + 2/3 of last part OR First 2 chars + last part	NA	NA
9	Adjacent	Exact	Exact or HES + 1	First part + 2/3 of last part OR First 2 chars + last part	Exact	Exact
10	Adjacent	Exact	Exact or HES + 1	First part + 2/3 of last part OR First 2 chars + last part	NA	NA
11	Exact or adjacent	Exact or Fuzzy	Exact or upto HES + 2	First part + 2/3 of last part OR First 2 chars + last part	Exact	Exact
12	Exact or adjacent	Exact or Fuzzy	Exact or upto HES + 2	First part + 2/3 of last part OR First 2 chars + last part	NA	NA
Only have first part postcode in STATS19						
13	Exact	Exact	Exact or HES + 1	Exact match on first part	Exact	Exact
14	Exact	Exact	Exact or HES + 1	Exact match on first part	NA	NA
15	Adjacent	Exact	Exact or HES + 1	Exact match on first part	Exact	Exact
16	Adjacent	Exact	Exact or HES + 1	Exact match on first part	NA	NA
17	Exact or adjacent	Exact or Fuzzy	Exact or upto HES + 2	Exact match on first part	Exact	Exact
18	Exact or adjacent	Exact or Fuzzy	Exact or upto HES + 2	Exact match on first part	NA	NA
Otherwise (i.e. no postcode information in STATS19)						
19	Exact	Exact	Exact	Not available	Exact	Exact
20	Exact	Exact	HES + 1	Not available	Exact	Exact
21	Adjacent	Exact	Exact	Not available	Exact	Exact
22	Adjacent	Exact	HES + 1	Not available	Exact	Exact
23	Exact	Fuzzy	Exact	Not available	Exact	Exact
24	Exact	Fuzzy	HES + 1	Not available	Exact	Exact
99	Other cases - linkage rejected as incorrect					

* A combined variable with 7 groups based on casualty class and casualty type (see text)

3.3 Results of linkage

3.3.1 Linkages by level and year

The number of linkages made at each level is shown in Table 3.2. In total, over 190 thousand out of nearly 600 thousand HES records were linked to STATS19 for the years 1999 to 2009, representing a rate of around one-third (32%). As can be seen from the table the proportion of HES records linked to STATS19 remains broadly stable over time, but is lower in the most recent two years (2008 and 2009)[12].

Agreement levels 1 and 2 account for over half of the links made. The number of links at these levels increases over time, and this appears to be largely related to the availability of casualty postcode on the STATS19 file, which is also increasing over this period. Conversely, the proportion of linked records at levels 19 to 24 (where postcode is not involved) falls between 1999 and 2009.

Table 3.2 Number of linked records by agreement level and HES year of admission

Level	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	1999-2009	% of total links
1	2,779	3,915	4,234	4,240	4,369	5,210	6,261	6,687	7,157	6,685	6,939	**58,476**	*30.5*
2	2,443	2,915	3,161	3,131	3,431	3,884	4,713	4,881	5,377	4,806	5,160	**43,902**	*22.9*
3	208	249	269	269	316	347	478	497	514	493	565	**4,205**	*2.2*
4	354	428	417	447	493	565	761	758	736	725	760	**6,444**	*3.4*
5	267	347	403	416	414	446	581	576	590	593	658	**5,291**	*2.8*
6	330	397	403	417	441	557	634	651	599	567	597	**5,593**	*2.9*
7	629	741	691	674	678	688	942	925	871	804	814	**8,457**	*4.4*
8	572	552	596	544	537	555	741	728	711	656	643	**6,835**	*3.6*
9	43	55	50	53	52	47	73	59	74	61	61	**628**	*0.3*
10	70	92	77	85	76	81	111	120	123	88	116	**1,039**	*0.5*
11	69	85	86	75	72	70	106	110	85	88	100	**946**	*0.5*
12	97	91	109	73	97	83	135	122	125	125	153	**1,210**	*0.6*
13	434	535	543	456	415	397	193	198	114	225	166	**3,676**	*1.9*
14	489	422	431	381	305	250	121	95	68	124	81	**2,767**	*1.4*
15	59	42	47	47	41	25	21	16	13	20	14	**345**	*0.2*
16	84	71	80	58	103	62	22	20	16	15	27	**558**	*0.3*
17	58	67	58	48	42	50	40	33	17	22	34	**469**	*0.2*
18	121	127	90	79	79	62	35	30	15	19	18	**675**	*0.4*
19	4,004	3,503	3,218	2,958	2,958	2,573	1,588	1,074	966	814	673	**24,329**	*12.7*
20	1,236	1,257	1,218	1,185	1,125	953	524	379	350	319	254	**8,800**	*4.6*
21	266	252	252	235	223	217	120	75	74	64	43	**1,821**	*1.0*
22	115	102	120	112	99	100	56	40	20	34	21	**819**	*0.4*
23	405	393	347	306	403	267	197	124	107	94	92	**2,735**	*1.4*
24	187	191	172	186	215	143	98	78	56	52	44	**1,422**	*0.7*
Total links	15,319	16,829	17,072	16,475	16,984	17,632	18,551	18,276	18,778	17,493	18,033	**191,442**	*100*
HES records	52,347	51,093	50,078	49,580	53,023	54,339	57,876	56,826	56,537	56,246	59,955	**597,900**	
% of HES records linked	*29*	*33*	*34*	*33*	*32*	*32*	*32*	*32*	*33*	*31*	*30*	***32***	

[12] An analysis suggests that this is due in part to an increase in the number of HES records which relate to non-traffic accidents (it is seen in section 4.1 that these are considerably less likely to be linked).

Table 3.3 shows the number of serious and slight STATS19 records linked to HES, according to whether or not a valid postcode is present in the police data[13]. It can be seen that a higher proportion are linked when postcode is available, and thus increasing availability of postcode in STATS19 is one reason for the increasing proportion of STATS19 records linked over time. However, considering STATS19 records with and without valid postcodes separately, in both cases there is an increase in the proportion linked to hospital records over time. This probably reflects an increasing tendency for road casualties to be admitted to hospital, and will be explored in more detail later (see section 5).

Table 3.3: Linked records as a proportion of STATS19 records by police severity, postcode validity and STATS19 accident year, England: 1999-2009[2]

	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
STATS19 serious records	**33,710**	**32,951**	**32,176**	**31,285**	**29,292**	**27,057**	**25,210**	**24,856**	**24,218**	**22,246**	**21,326**
No valid postcode	21,513	17,973	16,169	15,472	14,157	10,569	6,125	5,243	4,165	3,661	2,967
Valid postcode	12,197	14,978	16,007	15,813	15,135	16,488	19,085	19,613	20,053	18,585	18,359
Matched serious	**9,267**	**10,182**	**10,305**	**10,018**	**9,918**	**10,038**	**10,321**	**10,308**	**10,821**	**9,981**	**10,067**
No valid postcode	4,832	4,505	4,229	3,892	3,754	2,912	1,711	1,443	1,236	1,111	926
Valid postcode	4,435	5,677	6,076	6,126	6,164	7,126	8,610	8,865	9,585	8,870	9,141
Matching rate	***27***	***31***	***32***	***32***	***34***	***37***	***41***	***41***	***45***	***45***	***47***
No valid postcode	*22*	*25*	*26*	*25*	*27*	*28*	*28*	*28*	*30*	*30*	*31*
Valid postcode	*36*	*38*	*38*	*39*	*41*	*43*	*45*	*45*	*48*	*48*	*50*
STATS19 slight records	**248,494**	**249,855**	**244,586**	**234,755**	**225,603**	**218,991**	**212,539**	**201,026**	**192,744**	**179,788**	**173,574**
No valid postcode	147,534	122,475	108,316	103,232	98,967	77,000	42,989	34,901	28,380	29,234	22,690
Valid postcode	100,960	127,380	136,270	131,523	126,636	141,991	169,550	166,125	164,364	150,554	150,884
Matched slight	**6,059**	**6,643**	**6,770**	**6,461**	**7,070**	**7,585**	**8,233**	**7,967**	**7,954**	**7,516**	**7,967**
No valid postcode	3,335	3,120	2,973	2,746	2,869	2,491	1,446	1,176	984	965	825
Valid postcode	2,724	3,523	3,797	3,715	4,201	5,094	6,787	6,791	6,970	6,551	7,142
Matching rate	***2***	***3***	***3***	***3***	***3***	***3***	***4***	***4***	***4***	***4***	***5***
No valid postcode	*2*	*3*	*3*	*3*	*3*	*3*	*3*	*3*	*3*	*3*	*4*
Valid postcode	*3*	*3*	*3*	*3*	*3*	*4*	*4*	*4*	*4*	*4*	*5*

1: The validity marker used here is slightly different to the variable used in the matching process. However, the differences were small and the overall conclusions are unaffected.
2: There may be discrepancies between the year of hospital admission and accident year since injuries may not be immediately obvious and there may be a delayed between accident date and hospital admissions. In particular, there were nine reported casualties in recorded in 2009 in STATS19, but matched to 2010 HES records.

Tables 3.2 and 3.3 present initial results of linkage by year. However, it is important remember that the improvements in recording of postcodes is an artefact of the linking process which should be taken account of before carrying out trend analysis. The method used to adjust the number of linked records to take this artefact into account is described in Section 5.

Cross-sectional analysis of the proportion of records linked, according to different variables in the police and hospital datasets are explored in Section 4. In these analyses, adjustment described above is less important and so were not carried out.

[13] By definition, all casualties in STATS19 admitted to hospital should be coded as seriously injured, so cases where a link is made to a slightly injured casualty represent misclassification of injury severity by police. This is explored further in section 4.

3.3.2 The enhanced file

After the linkage process was completed, we created an enhanced file which supplemented the STATS19 details recorded by the police with medical details recorded in the HES file. This file includes many of the standard variables available in the HES system relating to admission and discharge information from hospital (e.g. dates and methods of admission and discharge), clinical information relating to the admission (e.g. diagnosis codes, including the external cause codes), episode and spell information, geographical information and patient information (e.g. age and gender of the patient).

In addition, the other variables can be derived from the diagnosis codes as mentioned previously in section 2.2. Finally, it is also possible to derived variables relating to severity of the injuries, for example, the length of stay in hospital and MAIS, the Maximum Abbreviated Injury Scale value.

The length of stay variable counts the number of nights spent in hospital between admission and discharge for the patients' spell in hospital, with a length of stay of 0 representing a patient admitted to and leaving hospital on the same day. Where a patient has more than one episode of care, this will be missing so the information available in the linked data file will probably tend to underestimate the true duration of treatment required, as those admitted for multi-episode spells are likely to be admitted for longer on average.

The Abbreviated Injury Scale (AIS)[14] is an internationally recognised method of measuring injury severity, developed by a committee of specialists for use in crash investigation work. The AIS is based on threat to life but also takes account of permanent impairment resulting from the injury and the energy dissipation required to cause the injury. The Scale is shown in Table 3.4. The body is divided into 6 regions and an AIS score assigned to each region. The MAIS for a casualty is the maximum of the AIS scores assigned and is used to summarise overall injury severity.

Table 3.4: The Abbreviated Injury Scale (AIS)

Code	Injury severity
AIS 0	No injury
AIS 1	Minor injury
AIS 2	Moderate injury
AIS 3	Serious injury
AIS 4	Severe injury
AIS 5	Critical injury
AIS 6	Maximum injury
Additional codes assigned in STATS19-HES file:	
AIS 9	Unknown
AIS 99	Cannot be coded (no suitable injury diagnosis recorded)

For the linked STATS19-HES data, AIS scores using the 1998 revision were estimated from ICD-10 coding of injury diagnoses using the mapping developed at the University of Navarra for the Apollo project [9].

Some limitations of AIS should be noted:

[14] http://www.aaam1.org/ais/

- AIS on its own is unable to predict mortality or outcomes. Scores 5 and 6 represent the "threat to life" associated with an injury and are not intended to provide a comprehensive measure of severity.
- AIS is not a true scale (e.g. the difference between AIS1 and 2 is not the same as between AIS4 and 5).

In addition to the general limitations of using MAIS as a measure of injury severity, there are some more specific issues relating to assignment of MAIS to the linked STATS19-HES data. For example:

- Not all HES records contain an injury diagnosis (i.e. a code in chapters S or T of the ICD-10 classification, see table 2.5). This may be due to inaccuracies in coding by hospitals, or possibly that some casualties admitted following road accidents were not injured. Such records are assigned a MAIS of 99, to distinguish from cases where an injury is present but of unknown severity.
- Not all injury diagnoses could be assigned an MAIS score; there are some S and T codes which do not appear in the mapping used. These are also assigned MAIS 99.
- For largely practical reasons, assignment of MAIS was based on the first 6 diagnoses codes, out of a possible 20 diagnoses recorded on HES for most of the years considered. Section 2 suggests that the proportion of HES records with more than 6 diagnoses represents around 1% of the total records, so this is unlikely to have a large impact, although it is possible that having more injuries coding these will be on average more severely injured.

It is likely that the combined effect of the above factors would understate the severity of road casualties appearing in the hospital data, which places some limitations on the comparability of the linked data with similar studies. However, *if* it can be assumed that the coding of injury by hospitals is broadly consistent over time, this is not a necessarily a problem in looking at data for England in isolation.

3.4 Assessing quality of linkage

It is important to attempt to assess the quality of the links made between the STATS19 and HES datasets, in order to understand the robustness of the conclusions that can be made based on the results. This is not straightforward, as there is a lack of strong identifiers which can be used to confirm a linked pair as a true match corresponding to the same casualty (for example, names and addresses are not available on either dataset).

As discussed above, postcode has relatively high power to discriminate, but is not always available and as it is used as a variable in the linkage process cannot easily be used to review quality of links. Some manual review may be possible (for example, to assess whether clear recording errors in postcode, such as transposed digits). However, this is time consuming and was therefore limited for this study. In summary, it is very difficult, if not impossible, to assess the quality of linkage precisely.

As a result we rely on other indicators to assess the quality of the linkages made, both false positives (cases where a link has been made but the casualties are non-matching) and false negatives (missed matched - cases where matches representing the same casualty but not linked). The following summarises the approach and resulting estimates. Further details are also given in Annex A.

3.4.1 False positives

To assess the number of false positives an empirical method was used based on the approach used in a previous Dutch study [14]. This method involved artificially inflating the 2004 STATS19 file with all STATS19 records from 2003 and 2005, with the years changed to 2004. The records from 2003 and 2005 were clearly false linkages, and by comparing the proportion of hospital records linked to the data from these other years we can estimate the likely proportion of false positives.

The results of this analysis suggest that overall the proportion of links that are false positives is likely to be around 3%, with some variation by agreement level. In particular, where there is exact agreement on postcode the false positive rate is estimated to be less than 1%. See Annex A for further details.

3.4.2 False negatives

To assess false negative rates (i.e. missed matches), a broad assessment was made based on a probabilistic calculation, essentially a simplified version of the approach used in French studies [16]. For details, please see Annex A.

As a proportion of linked records, the overall estimated number of false negatives for the 2004 data is around 13%. However, this depends on whether a valid postcode is available in STATS19. The availability of postcode provides strong power to discriminate and allows more tolerance in other matching variables used (without introducing an unacceptable number of false positives). The false negative rate is estimated to be around 5% for postcoded records compared with nearly 50% of non-postcoded cases (Table A2). As the proportion of STATS19 records with valid postcode has increased over the period for which data have been linked, the proportion of missed matches is likely to have fallen between 1999 and 2009. This needs to be considered for in any analysis of the number of links achieved over time.

3.4.3 Missing or invalid data

The above calculation of false negatives is based on an empirical approach, which allows for some errors in data (for example, estimation of age by the recording police officer), but not for cases where data is completely missing for the four key matching variables (age, sex, region and date). Annex A illustrates the broad proportion of records on each of the data files with missing data and suggests that around 4% of matches may have been missed in addition to the false negatives estimated above. This should be kept in mind in interpretation of results.

3.4.4 Potential Bias from the matching

There may be potential biases arising from the matching process itself. As already discussed, the completeness of postcode information improves over time, leading to a bias towards later calendar periods. In addition, the post code completeness varies by police force, thus the matching would be biased towards regions with higher completeness of postcode data.

The matching bias may potentially be explored by looking at the variation of matching rate by different characteristics. However, there are a few issues to consider prior to this.

First, neither dataset was completely matched, so we need to consider how the matching bias may affect both datasets: does the matching rate to the HES data vary by different characteristics of the police-recorded casualties, and whether the matching rate to the STATS19 data vary by different characteristics of the hospital admitted patients.

Second, there may be known or suspected reasons for the variation in the matching rate, which are independent of any bias due to the matching process itself.

There may be variations in the matching rate to the police data for different road user group recorded in the HES dataset. However, this could be due to differential reporting rate to the police for different road user groups. For example, it is well known that single vehicle pedal cyclists accidents are less likely to be reported to the police compared to other type of accidents.

More seriously injured casualties would be more likely to be admitted to hospital. So it would be unusual if the matching rate to the HES dataset were the same for different injury severity recorded in STATS19. In addition, variations in the hospitals practices by certain characteristics could also lead to differences in the rate of matching to the HES dataset. For example, children or the elderly may be more likely to be admitted to hospitals as a precaution, thus creating differences in the matching rate to the HES records by different age groups recorded in STATS19.

The variations in the two types of matching rate will be explored in Section 4. However, since these variations will not only reflect matching bias but also various other external factors, it is would be difficult to comment on the matching bias using these matching rates alone.

It is likely external factors that affect matching rates will have a greater impact than those due to matching bias, masking any effects due to matching bias. Thus, this makes it difficult to assess any matching bias through studying variations in the matching rates. However, it is likely that any unexpected or unexplainable patterns found during the analysis could be potentially due the matching process (see Section 4 for further details).

3.5 Discussion and conclusion

In this section, we have outlined the method used to link STATS19 and HES datasets, and briefly presented an assessment of the quality of linkage achieved.

Ultimately there is trade off between false negatives and false positives, maximising the former is likely to increase the latter, and conversely minimising the latter will tend to reduce the former. Our approach can generally be considered conservative, with emphasis on a minimal number of false positives. This produces a linked dataset for analysis which contains a considerable number of sufficiently robust links for analysis, but requires some adjustment for missed matches (and some assumptions) in order to estimate the likely proportion of true matches between police and hospital datasets that could be achieved.

Adopting a probabilistic matching methodology, as has been used in a number of other EU countries, may provide a more formal basis for assessing the quality of the linkage process. However, this would require considerable efforts to develop and was not carried out in this case. For now we can conclude that the overall quality of linkage achieved is sufficient for broad analyses of the linked data, but may not be suitable for trends analyses due to matching bias.

There are biases from the matching process. However, most of these may be hard to quantify as these effects are likely to be masked by variations in the matching rate due to external factors. It is likely any potential bias from the matching process will present as unusual/unexplainable patterns discovered as part of analysis of the matched dataset. The DfT welcomes any feedback on matching biases from analysts who have used the matched data for analysis. Please email your comments to roadacc.stats@dft.gsi.gov.uk.

Overall, this section has demonstrated the validity of the approach to linking the two datasets, and its acceptable quality for the analyses. The dataset is likely to contain relatively few incorrectly linked records.

Compared with other studies using similar methods, the proportion of inpatient records linked appears to be conservative. Transport Research Laboratory (TRL) linked Scottish hospital and police data using a broadly similar rules-based approach [5,10]. The overall proportion of “road traffic accident” hospital inpatient records linked to STATS19 was 56% using the Scottish data [10], compared with the 41% achieved for HES records within the scope of STATS19, and the overall 32% achieved for HES records in this study. However, direct comparisons and firm conclusions are difficult due to differences in the identifier variables used for the matching. For example, the Scottish linkage did not use postcode as a linkage variable and may as a result contain a higher proportion of false positives. In addition, there are differences in the hospital data sources, for example, the Scottish hospital dataset has also been available for considerably longer than the equivalent HES data for England and may identify road casualty admissions more accurately. Further, hospital practices may differ between Scottish and English hospitals, for example, on the use of short stay ‘observation’ wards.

There have been many other studies comparing police and hospital data on road casualties over the last few decades. A study comparing police casualty data with hospital road accident inpatient records for part of Scotland [19] linked around half of inpatient records to STATS19 killed or seriously injured casualties. In this study, the authors were able to use the patient level data, rather than derived patient level data (from episodes) as in this study. So it is possible that our duplication of the episode data may not completely eliminate multiple episodes to a single patient, and thereby deflating the matching rate.

Another study covering 16 A&E departments across Britain in 1993 [12], found that 62% of patients were linked to police records. Conversely, 49% of police serious and 6% of police slight casualties were admitted as inpatients. These results are again suggestive of lower than expected proportions of the linked STATS19-HES records found in this study. However, the 1993 study was from an earlier time period compared to the current study, and there may be differences in recording methods and other practices may have changed over time. For example, the use of matching by casualty postcode has only been possible since police started recording this in 1999.

There are a number of possible reasons for this lower proportion of linked records, including:

- Cases where a casualty was admitted to hospital but does not appear in the HES extract used for linking. For example, work in the Netherlands found that around 17% of police serious casualties were linked to records not coded as road accidents in the inpatient data [14]. Here, the HES file includes non-traffic (off road) accidents, but not those recorded as falls or where no external cause of injury is recorded.
- Past analyses have suggested that up to 10% of all HES injury records (or patients with an “Injury, poisoning and certain other consequences of external causes” diagnosis code) do not have a code for the cause of injury (or a “External causes of morbidity and mortality” code recorded). Such an estimate suggests this group is of the same order of magnitude as the total number of records recorded as transport accidents [20].
- Cases where links were made within an acceptable level of tolerance, but there was more than one possible match and it was not possible to distinguish between the two with the information available (‘unresolved duplicates’).
- Missing data may prevent links being made (a crude estimate suggests this may result in potentially 4% of correct linkages being missed)
- The defined tolerances allowed for two linked records to be considered as a match may be too strict, resulting in missed matches (although as noted above, if relaxed this would result in more incorrect linkages being accepted as matching).

The above should be borne in mind when interpreting the following analysis. However, given the uncertainties it seems difficult to quantify the extent of these effects without detailed further study. It is possible that a formal probabilistic linkage method (as used in some other countries) might address the latter three points to some extent but would require considerable effort to develop.

4. Analysis of linked data

The following analysis explores how the proportion of records linked varies according to the values of particular variable in both the hospital (HES) and police (STATS19) datasets. The analysis considers key variables in turn. A more detailed approach using multivariable analysis to explore the effect of the factors after adjustment for other variables were not used in this study.

Throughout, the full linked dataset covering all years 1999 to 2009 is used. No attempts were made to adjust for missing matches due to the difficulty in imputing the characteristics of the missed matched. This assumes that the missed matches share broadly similar characteristics as the linkages made. In addition, there may be some false linkages included in the dataset; we assume at the aggregate level that these will not distort the conclusions drawn. However, as noted in section 3, it should be kept in mind that in general the achieved linkage rates are likely to be underestimates.

4.1 Propensity of hospital inpatients to appear in police data

Assuming the data linkage is broadly robust, factors associated with different linkage rates are likely to reflect factors associated with variations in reporting levels of accidents to the police (or subsequent recording levels once an accident has been reported to the police).

4.1.1 Type of accident

Table 4.1 shows the proportion of HES records linked to STATS19, according to the nature of the accident as recorded by the hospital[15].

Table 4.1 Linkage results by accident type recorded in HES, 1999-2009

Within the scope of STATS19[1]	**Accident Type**	**Linked to STATS19**	**Not linked to STATS19**	**Total HES records**	***Proportion linked***
Yes	**Total**	**176,153**	**257,359**	**433,512**	***41***
	Traffic	169,132	226,198	395,330	*43*
	Board/alight	726	12,643	13,369	*5*
	Unspecified	6,295	18,518	24,813	*25*
No	**Total**	**13,569**	**144,093**	**157,662**	***9***
	Traffic	88	270	358	*25*
	Non-traffic	13,221	110,160	123,381	*11*
	Board/alight	39	1,753	1,792	*2*
	Unspecified or not known	221	31,910	32,131	*1*
Unknown		**1,720**	**5,006**	**6,726**	***26***
All		**191,442**	**406,458**	**597,900**	***32***

1 Defined as an accident type which should be reported by police in STATS19, as set out in the document STATS20 which contains instructions for completing STATS19 [16]

[15] This is derived from the ICD-10 code identifying cause of injury, via a lookup table produced by DfT to identify which codes relate to accidents within the scope of the STATS19 definition of a road accident [15].

It is clear that a higher proportion of casualties admitted in accidents within the scope of STATS19 are linked to the police data (41% compared with 32% overall). In particular 43% of cases recorded as traffic accidents are linked. Allowing for the inaccuracies in the linkage procedure would probably result in a proportion closer to half.

Over a third of the unlinked HES records are those recorded as non-traffic accidents or accidents occurring off the public highway. If these are coded correctly we would not expect to find them within STATS19. In practice, there will be some degree of miscoding by the hospital, as is evident from the fact that around 9% of such records are linked to STATS19[16].

The remainder of the analysis in section 4.1 will focus on HES records coded as traffic accidents within the scope of STATS19 (395,330 records) which are of more interest when exploring variations in completeness of police data. Table B1 (Annex B) shows that within this group the proportion of HES records linked to STATS19 varies according to the nature of the collision recorded. A lower proportion of inpatient casualties in non-collision accidents were linked to STATS19 compared to collision accidents[17].

4.1.2 Road user type

Figure 4.1 shows how the proportion of HES records linked to STATS19 varies by road user type, for the main road user groups and broad nature of accident[18].

Overall, the linkage level was the highest for pedestrians followed by car users, with lowest for bus occupants and pedal cyclists (green bars). However, if non-collision accidents are removed then the proportion of pedal cyclists admitted that are linked to STATS19 is broadly in line with other road users (blue bars). In collisions recorded as being with a car or van, the linkage rates for vulnerable road users are highest and are similar across the groups (pedestrians 58%, pedal cyclists 59% and motorcyclists 62%; light-blue bars).

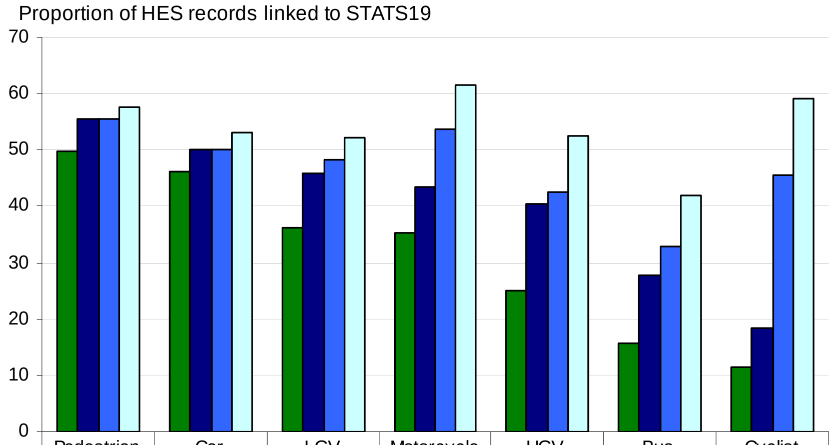
Table B2 in Annex B shows the number of linked and unlinked records by road user type in non-collision accidents. Nearly half of these involve pedal cyclists, of which only 3% are linked to STATS19. These 43 thousand unlinked non-collision pedal cycle records account for nearly a fifth of the unlinked HES traffic accident records. If these were removed, the proportion of traffic accident inpatients linked to STATS19 rises to around 48%.

[16] It is likely that some of the unlinked non-traffic accident records will also represent genuine traffic accidents, that are miscoded in the HES data and not present in STATS19. It is also possible that within the HES traffic accidents are some incorrectly classified non-traffic accidents, so that the true proportion of casualties in traffic accidents known to police would be understated. [3] discusses this and suggests that pedal cycle accidents may be particularly affected.

[17] This is consistent with previous research (e.g. [3], [12])

[18] It should be noted that as road user group is used (as part of the road user class variable) in the linkage process, some caution may be needed in interpreting the results shown here.

Figure 4.1 Proportion of HES records linked to STATS19 by HES road user and accident type: 1999-2009



	Pedestrian	Car	LGV	Motorcycle	HGV	Bus	Cyclist
🟩 All	50	46	36	35	25	16	11
■ Traffic accidents only	55	50	46	44	41	28	18
🟦 Excluding non-collision or hit object	55	50	48	54	43	33	46
□ Collision w ith car/LGV	58	53	52	62	52	42	59

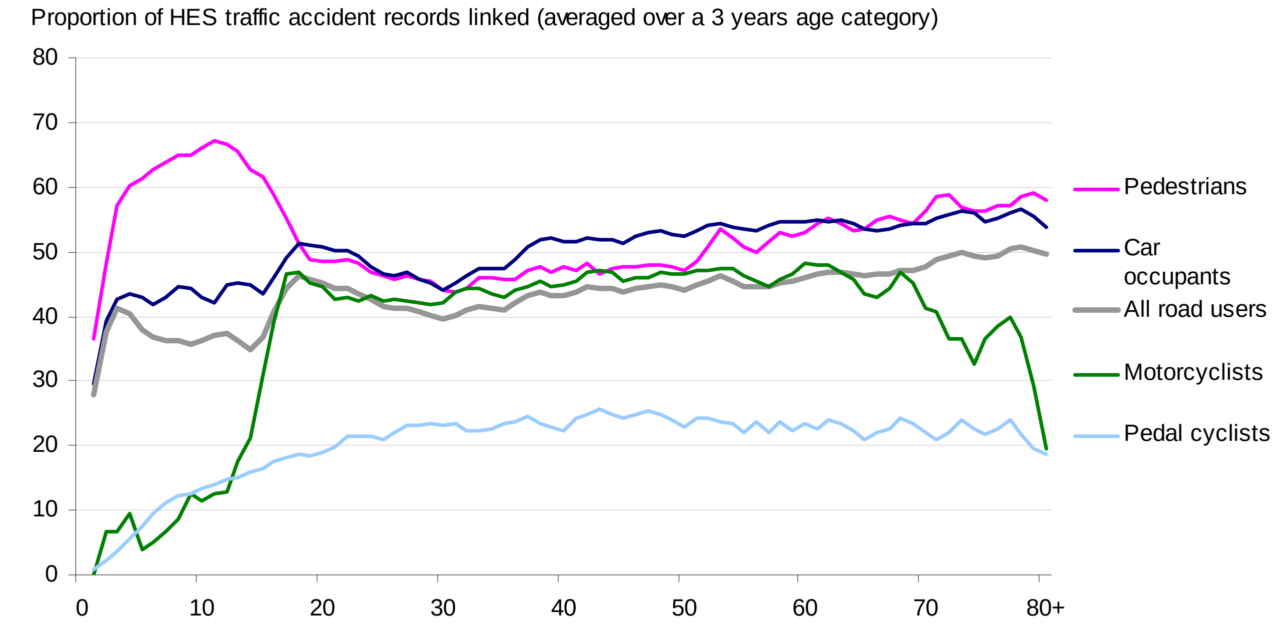
4.1.3 Age and gender

Figure 4.2 shows how the proportion of HES traffic accident records linked to STATS19 varies by age for the four main road user groups. The highest proportion of records linked are for child pedestrians, and the lowest for child cyclists and motorcyclists. This is likely to be partly because the number of child cyclists and motorcyclists are small, and so even a small absolute number of unlinked records may appear to be a high proportion of all casualties. In addition, these lower linkage rates may also reflect a relatively high number of non-collision accidents in these road user groups.

For ages between 16 and 70 the proportion of car occupant, motorcyclist and pedestrian casualties linked is very broadly similar, around half of HES records. Linkage rates for pedal cyclists are however considerably lower across all age groups.

There are no clear relationships between linking rates and gender. Overall, a marginally higher proportion of female casualties were linked. However, this may be an artefact of the variation in gender distributions by road user groups, which is linked to linkage rates as discussed above. For example, a higher proportion of male casualties admitted to hospital are pedal cyclists, and pedal cyclists are less likely to be linked to STATS19 records (Figure 4.1).

Figure 4.2 Proportion of HES traffic accident records linked to STATS19 by age and road user type recorded in HES, 1999-2009

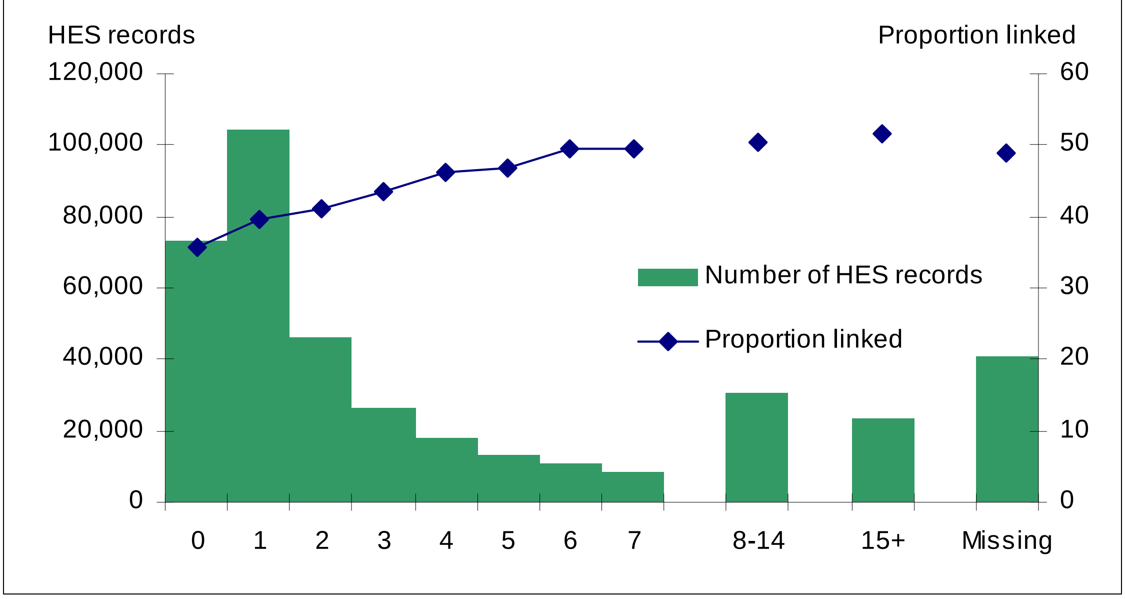


4.1.4 Casualty severity measures and diagnoses

The proportion of HES records linked to STATS19 increases with length of stay in hospital, which might be considered as a crude proxy for severity of injury (Figure 4.3). Around half of those spending a week or more in hospital are linked to STATS19, compared with 36% of those admitted and discharged on the same day.

The Maximum Abbreviated Injury Scale (MAIS, see section 3.3.2) can also be used as a measure of severity of injury, and similarly the linkage rate is higher for those records with MAIS 3+ (50%) compared with MAIS 1 or 2 (42%) (Table B4, Annex B).

Figure 4.3 Proportion of HES traffic accident records linked to STATS19 by length of spell in hospital (days), 1999-2009

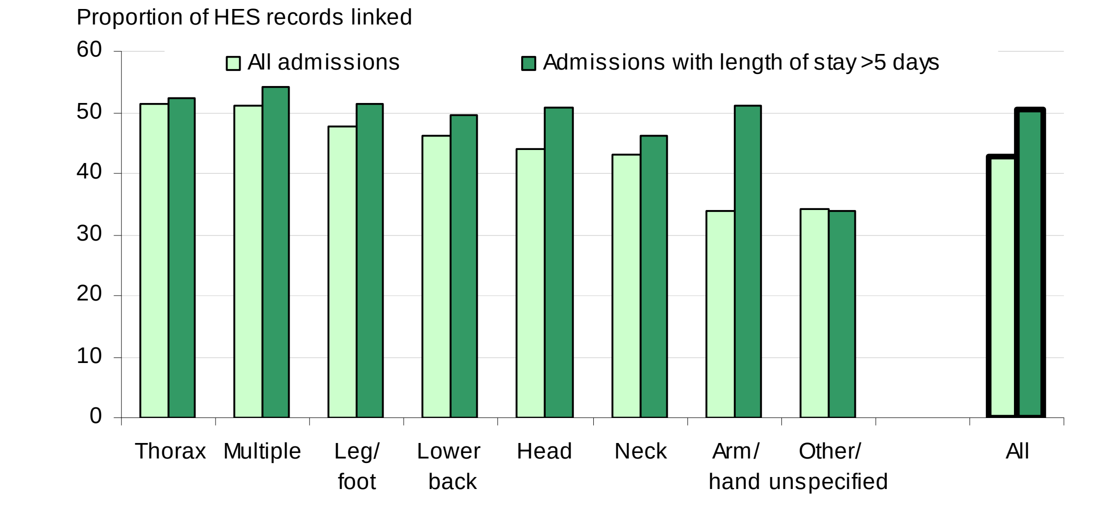


Based on the primary diagnosis recorded in HES, lower proportions of casualties with ‘dislocations, sprains or strains’ are linked to STATS19 (37%, see Table B5, Annex B). This lower proportion may be due to the difficulties of assessing the severities of these injuries at the roadside by the police, or due to perceptions that such injuries are unlikely to require admissions to hospitals, leading to miscoding of these casualties as uninjured[19]. Around half of the linked and unlinked records in the HES dataset had a primary diagnosis of fracture.

Figure 4.4 shows variations in linkage rates for HES records by body region of primary injury. Variations here will to some extent reflect the types of injury associated with the different parts of the body, for example, neck injuries are more likely to be dislocations, sprains or strains compared to other body regions.

[19] This is a crude analysis based on one of (as many as) 20 possible diagnoses and could benefit from more detailed work could investigates. However, any analysis will be limited by the imperfect nature of the diagnosis coding in the HES file [3].

Figure 4.4 Proportion of HES traffic accident records linked to STATS19 by body region of injury of primary diagnosis, 1999-2009

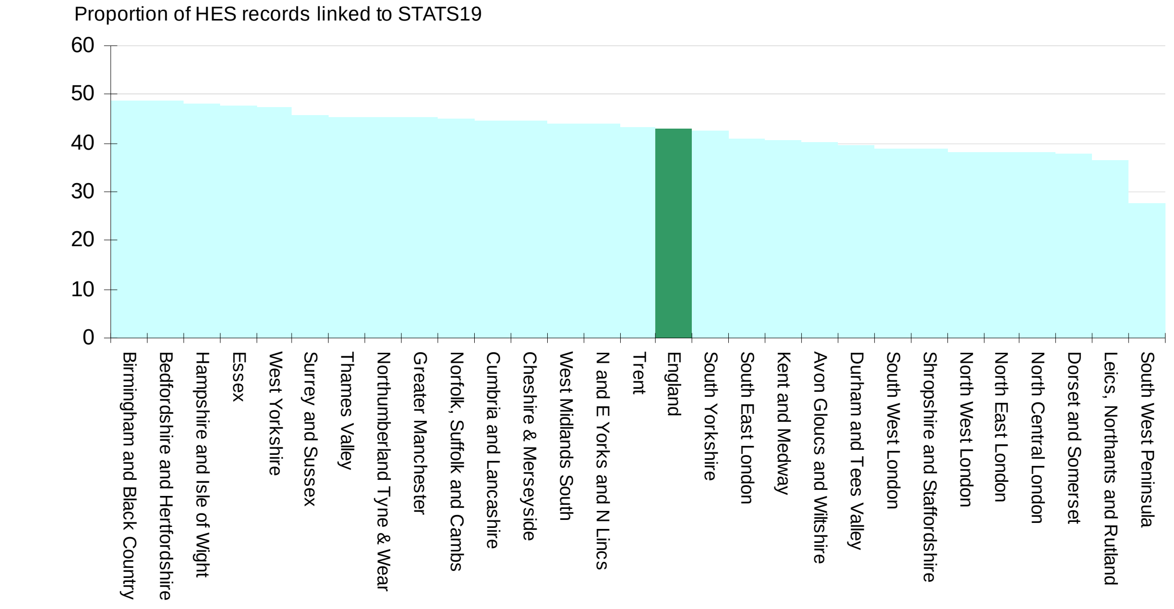


4.1.5 Strategic Health Authority

Geographic variations are difficult to interpret as the quality of linking variables, particularly recording of postcode in STATS19, varies by police force area. Figure 4.5 shows no obvious pattern in linkage rates by Strategic Health Authority (SHA), with only two areas having a matching rate that deviate more than 6 percentage points away from the average for England[20].

[20] One of these areas, South West peninsula, is covered by the Devon and Cornwall police force which has a particularly low proportion of records with valid postcodes in STATS19 for the years linked.

Figure 4.5 Proportion of HES traffic accident records linked to STATS19 by SHA, 1999-2009



4.1.7 Summary and discussion

The above analysis suggests that at the national level and for the more severely injured casualties (represented by admissions to hospital), STATS19 appears to be broadly representative of the casualty population:

- Overall, 41% of accidents recorded in HES and coded as within scope of the STATS19 definition of a road accident are linked to STATS19. This proportion rises to 48% for traffic accidents excluding non-collision pedal cycle accidents. Given the likely underestimation of the number of records linked (due to missed matches), this suggests that over half of those admitted to hospital are recorded in STATS19. This would be broadly consistent with previous studies (e.g. [3]).
- Excluding non-collision pedal cycle accidents, the proportion of admitted casualties appearing in the police data is broadly similar across the main road user groups, although bus occupant casualties may be under-represented in STATS19.
- A very low proportion of pedal cycle casualties admitted to hospital following a non-collision accident become known to police (even when it is recorded as on a public road), but in collision accidents the proportion of pedal cyclist admissions appearing in STATS19 is comparable with other road user groups.
- Based on this analysis, the likelihood of linkage between HES and STATS19 records are broadly similar between different age and genders. However there is some evidence of higher linkage rates for child pedestrian casualties.

- Within casualties admitted to hospital, those with more severe injuries (as indicated by a longer spell in hospital or by a high MAIS score) are more likely to be known to police.

The findings are generally in line with previous research. For example, analyses of the two datasets at aggregate level [1] concluded that a lower proportion of single vehicle pedal cycle and motorcycle accidents were likely to come to the attention of police. A recent linkage of police and inpatient data for West Scotland [19] found that non-collision accidents, adults (compared to children), females and shorter length of stay in hospital were factors associated with lower levels of reporting to police. In the most complete previous study [12], Simpson found lower proportions of hospital casualties (including A&E attendances) linked to police data for single vehicle accidents, (particularly for pedal cyclists, but also for other road user groups). In addition, Simpson found lower linkage proportions for slight compared to serious injuries (as measured by MAIS), but found no variation with gender. This study also found a lower proportion of casualties with whiplash injuries (sprains/strains to the neck area) in police compared with hospital data.

There are a number of other variables of interest that may be related to the propensity of an accident to become known to the police. However, this can only be analysed where both the linked and the unlinked HES records have the variable. Therefore variables of interest present only in the police file such as road type or speed limit cannot be analysed in this way in this study. Previous studies [12] have found that factors such as vehicle damage and method of transportation to hospital are among those most strongly associated with whether a casualty appears in police data, but this cannot be assessed using the results of this linkage.

There were no obvious unexpected results to suggest any biases from the matching process, although this does not mean there are no biases inherent within the process.

4.2 Propensity of casualties in police data to be admitted to hospital

Comparing linked and unlinked STATS19 records provides some insight into the propensity of casualties known to police to be admitted to hospital.

4.2.1 Severity recorded by police

As expected, the matching rate to the HES records is higher for seriously injured casualties compared to slightly injured casualties (as recorded in STATS19). In fact, in theory any casualty admitted to hospital as an inpatient should, by definition, be recorded in STATS19 as seriously injured. However, in practice 3% of casualties recorded as slightly injured by police were linked to HES (Table 4.2) and these linked cases accounted for around 40% of all linked records.

Table 4.2 Proportion of STATS records linked to HES by casualty severity recorded by police, accident year 1999-2009

Severity	Linked	Not linked	Total	*Proportion linked*
Serious	111,226	193,101	304,327	*37*
Slight	80,225	2,301,730	2,381,955	*3*
Total	**191,451***	**2,494,831**	**2,686,282**	***7***

* Please note that there were nine records for the 2009 STATS19 data that do not appear in the HES dataset until 2010. So there were a total of 191,451 records matched for STATS19 calendar period of 1999 and 2009. So there may be some discrepancies in the total number of linked figures between the following and previous sections.

4.2.2 Road user type and casualty class

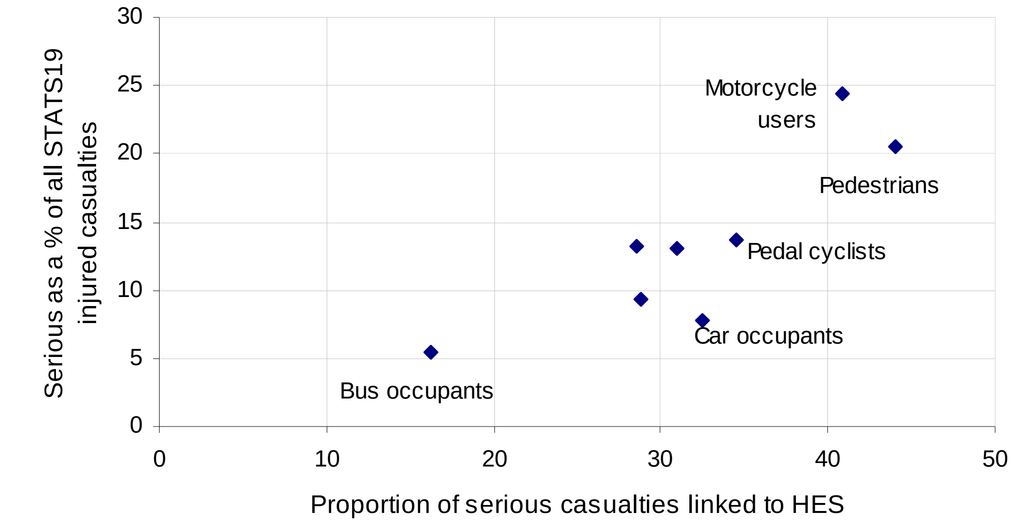
Pedestrians (44%) and motorcyclists (41%) have the highest proportion of seriously injured casualties linked to HES records (Table B6, Annex B), with the lowest rates for bus occupants (16%).

The matching variations between different road user types may be an artefact of the variations of severity distributions between the groups, where severity of injury is associated with the likelihood of being matched to HES as previously mentioned. Figure 4.6 shows a positive association between road users which have higher proportions of serious injuries in STATS19 and likelihood of being matched to HES records.

Drivers are more likely to be linked to HES records than passengers for all vehicle types except buses and coaches (Figure B1, Annex B). This pattern is consistent with the severity proportion (serious as a percentage of all injuries) shown in STATS19[21].

[21] It should be noted that casualty type and class are used as a variable in the linkage process, and therefore it is possible that these results depend on the (essentially subjective) linkage rules used. However, this is not expected to distort the overall conclusions shown here. This could be checked by basing the analysis on casualties with a valid postcode, where casualty class is less important as a linking variable.

Figure 4.6 Proportion of STATS records linked to HES by casualty type against STATS19 severity proportion (serious/all injuries)



4.2.3 Age and gender

For both seriously and slightly injured casualties recorded by the police, a higher proportion of child and elderly casualties are matched to the HES dataset (Figure 4.7). This may reflect the fact that accidents are more likely to have serious consequences for these age groups, or a tendency for hospitals to admit children and the elderly for precautionary reasons or observation[22].

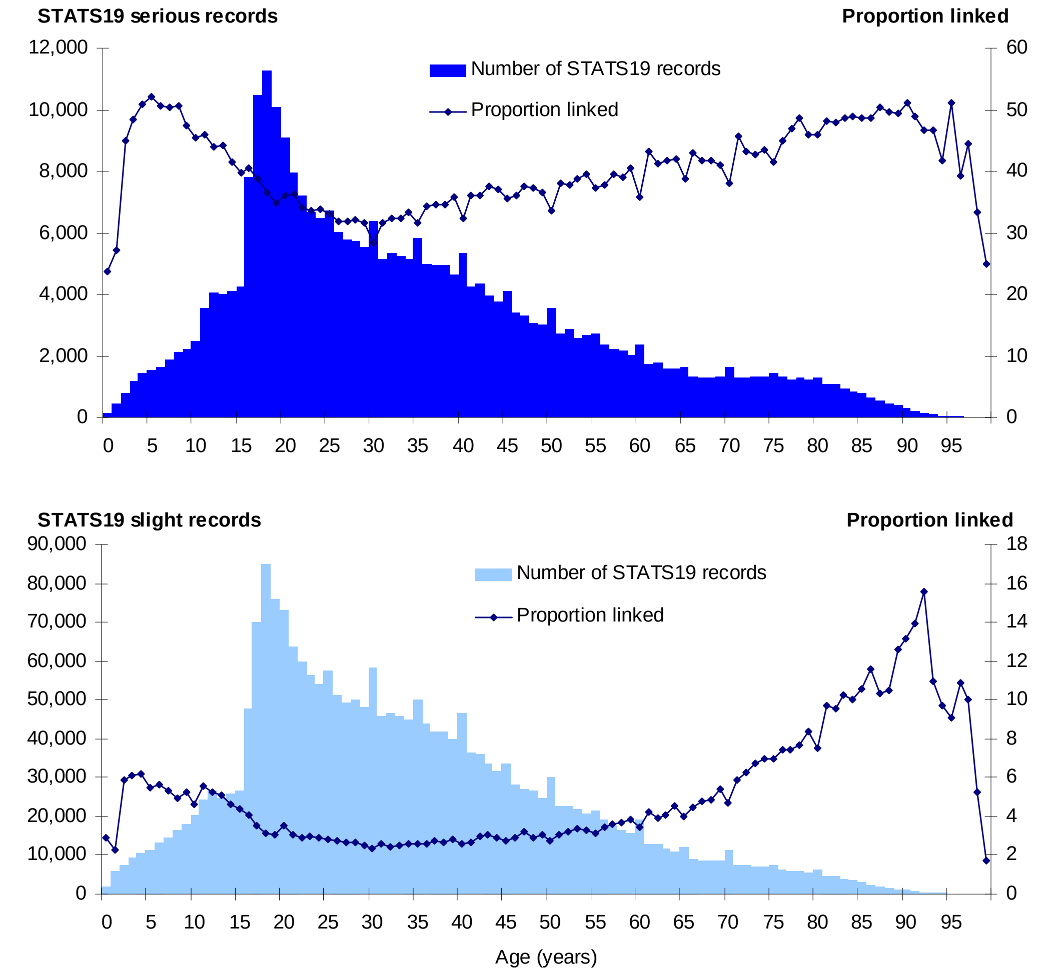
In addition, the number of reported casualties in STATS19 increased for individuals in their late teens and early twenties, but the matching rates to the hospital data were among the lowest for this age group (Figure 4.7). This may reflect the fact that individuals in this age group are less likely to seek medical treatment after a road accident or that hospital are less likely to admit individuals in this age group as they are (or perceived to be) healthier and fitter compared to other age groups.

Figure 4.7 also shows the effect of casualty age on linkage rate. Lower linkage rates occurred at some ages ending in zeros or fives (e.g. 30,35,40,etc). This may be due to the heaping of casualty ages estimated by the police as seen in Figure 3.2 in the previous section. So these variations in the linkage proportions are purely an artefact of the linkage process and the data quality, rather than true associations between these ages and the probability of linkage.

Overall, a higher proportion of male STATS19 serious casualties are linked to HES than female (38% compared with 34%; Figure B2, Annex B), which may reflect the fact that males are more likely to be involved in more severe road accidents

[22] Although this pattern will also reflect variations in type of road user by age group (in particular, a higher proportion of pedestrian casualties among the younger and older age groups). Further analysis, not presented here, suggests the age variation is present after allowing for variations in road user type.

Figure 4.7 Proportion of STATS serious casualties linked to HES by age, 1999 - 2009



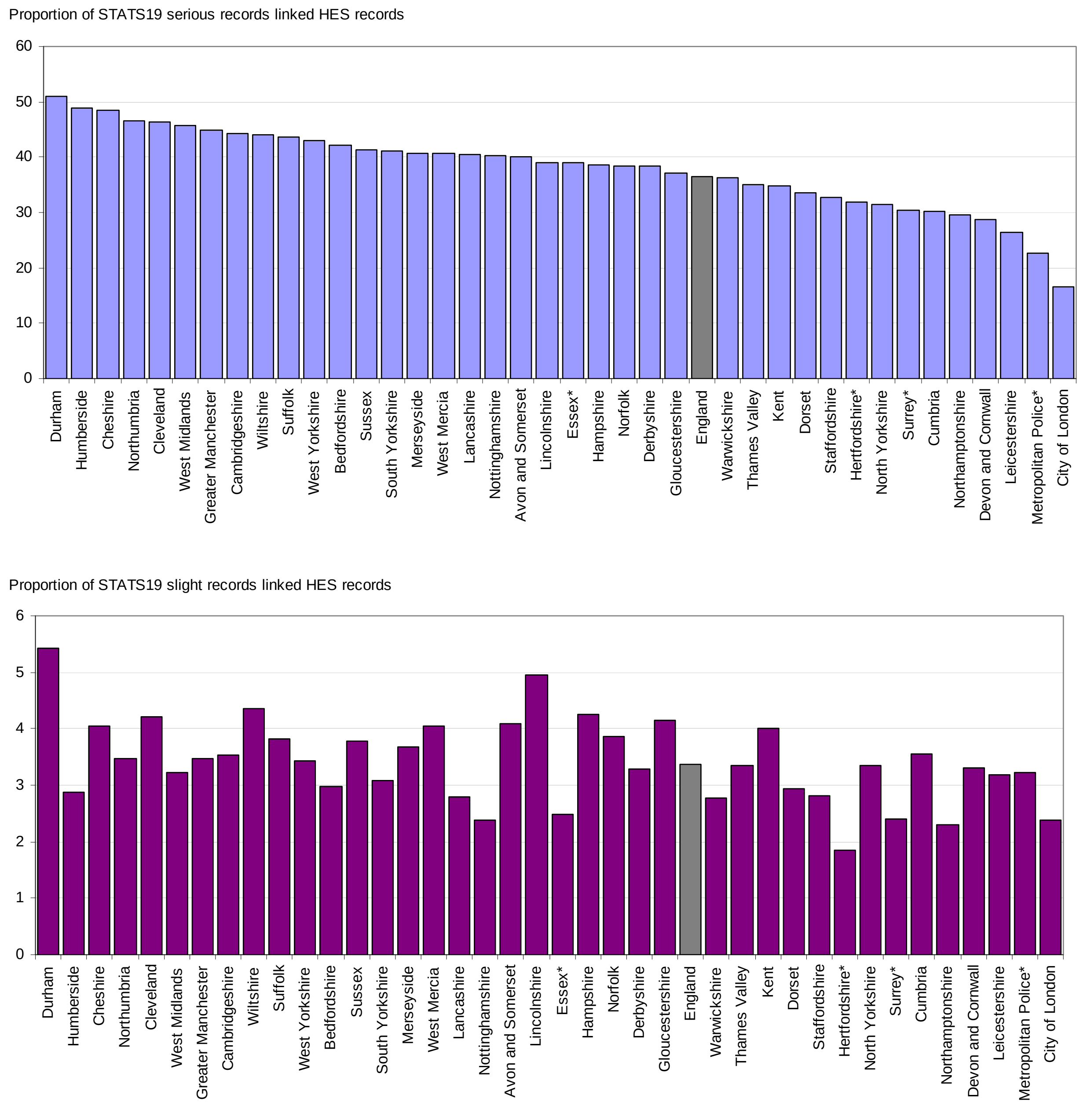
4.2.4 Police force area

In theory, the linked data offers an opportunity to assess how the proportion of serious casualties varies by police force, which could provide information related to any variations in coding of injury severity. However, in practice this is complicated by differences in the proportion of STATS19 records with a valid postcode across police forces, since the availability of postcode is associated with a better quality of linkage.

Figure 4.8 shows the proportion of STATS19 serious and slight casualties that are linked to HES for each police force in England. The England average is lower than median proportion of all the police forces, because some of the larger forces have lower proportions of linked casualties. In particular, the Metropolitan police (which has the lowest proportion) accounts for around one sixth of all STATS19 serious casualties in England over the period studied. However, this lower matching rate may be due higher volumes non-residents, tourists and commuters who may visit this area and are harder to match using the current methods.

The same ordering of forces is used for slight as for serious, so we can see that, although there is some association between the proportions of serious and of slight casualties linked, this is not particularly strong.

Figure 4.8 Proportion of STATS serious and slight casualties linked to HES by police force area, 1999 - 2009



* Until 2000, Metropolitan Police patrolled areas in Herts, Essex and Surrey. In these cases the local authority code relates to the area patrolled but the Police Force Code is "1" - Metropolitan.

Figure B3 (Annex B) shows the association between the proportion of STATS19 casualties linked, and the availability of postcode information. There is a weak positive correlation, which is stronger for serious casualties than slight with some notable outliers[23].

[23] For example, for serious casualties both City of London (55% of records postcoded but a linkage rate of 17%) and the Metropolitan police (57% of records postcoded but a linkage rate of 23%) appear to be outliers.

Taking these patterns together, there is evidence to suggest that availability of postcode has an influence on the patterns observed. However, this is perhaps not as strong as we might have been expected, particularly for slight casualties. This may indicate inaccuracies in the linkage process, or may be due to variations in the type of accidents occurring in different areas and in coding practices. A multivarible analysis may help to assess this further, but it is difficult to draw any firm conclusions from the results presented here.

4.2.5 Other variables

There are a number of other variables which may be of interest in assessing variations in propensity of STATS19 casualties to be linked to the HES dataset.

As an example, Table 4.3 shows that the propensity to be admitted to hospital (for those casualties appearing in STATS19) is higher where a police officer attends the accident scene compared with those reported elsewhere (e.g. at a police station). The former group accounts for the vast majority of STATS19 casualties, but again this may be an indication that a greater proportion of the more severe accidents become known to the police. This suggests that slight casualties are under-estimated relative to serious assuming generalisability to non-hospitalised casualties.

Table 4.3 Proportion of STATS19 casualties linked to HES by reporting method 1999 - 2009

	Serious			Slight		
Method of reporting	**Linked**	**Total**	***Proportion linked***	**Linked**	**Total**	***Proportion linked***
At scene	99,957	260,311	*38*	71,526	1,789,725	*4*
Elsewhere	7,482	29,470	*25*	6,252	479,954	*1*
Undefined	3,787	14,546	*26*	2,447	112,276	*2*
Total	**111,227**	**304,327**	***37***	**80,227**	**2,381,955**	***3***

4.2.6 Summary and discussion

This section has briefly looked at the proportion of STATS19 casualties linked to HES, which is assumed to be a measure of the propensity for road accident casualties to be admitted to hospital. In a crude sense, this might be considered as an *indication* of ‘more severe’ casualties within the serious category, although hospital admission is unlikely to be a perfect measure of this. This shows that among road casualties in the police data:

- Pedestrian casualties, followed by motorcycle users, are the road user groups most likely to be admitted as a result of their injuries.
- Children and elderly casualties are more likely to be admitted than other age groups.
- In general, the propensity for casualties to be admitted to hospital is correlated with higher proportions of injured casualties that are coded as serious in STATS19.

This is broadly consistent with other research. For example, past reports [e.g. 3] suggest that pedestrians are more likely to be admitted as a result of their injuries compared with car occupants, and that children are more likely to be detained for observation. It is also noted that the propensity of hospitals to admit (among those with less severe injuries) can also depend on socio-economic group and access to hospital, although these factors can not be analysed here.

There were no obvious unexpected results, although there are variations in the recording of postcodes by police force which may be considered as bias from the matching process. This in part explains some of the variation in the linkage rate to the HES records between police forces. However, there are notable outliers and further work is needed to explore the patterns shown.

4.3 Further analysis of linked data – casualty severity

This section gives details of the initial analysis of the linked dataset of some 190 thousand records to explore patterns of injury severity and factors associated with misclassification of injury severity by police. This includes looking at trends over time, based on proportions rather than absolute numbers, to allow for variations in the quality of linkage over time[24].

4.3.1 Police classification of injury severity

In STATS19, reporting police officers classify casualties as killed, seriously injured or slightly injured at the roadside, without extensive medical knowledge. The definitions of serious and slight injuries are:

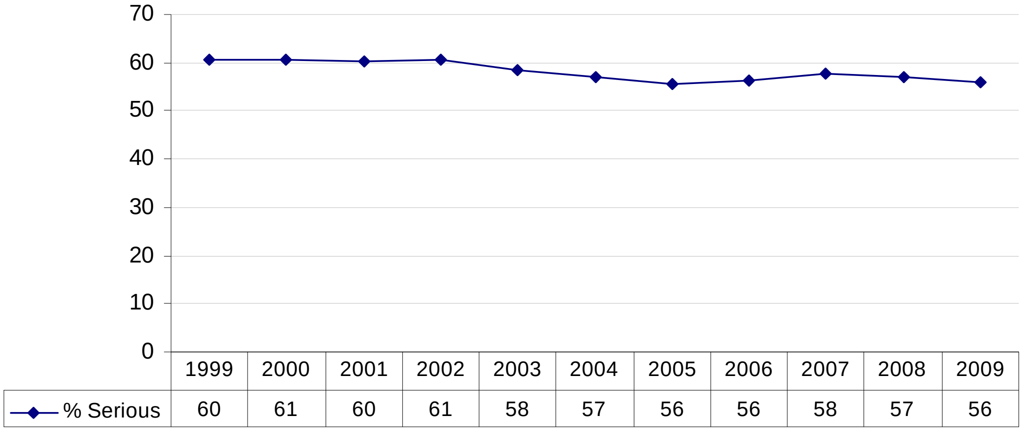
- **Seriously injured:** those *detained in hospital as an inpatient*, or any with particular types of injury, regardless of whether they are detained, including fractures, concussion, internal injuries, crushings, burns, severe cuts and severe shock.
- **Slightly injured:** those with other injuries, for example sprains (including neck whiplash injury), bruises and cuts which are not judged to be severe or slight shock requiring roadside attention.

Therefore in theory any casualty appearing in the linked dataset should be recorded as seriously injured in STATS19, and linked records coded as slight are assumed to represent casualties miscoded by police[25]. Overall, around 58% of linked records are coded serious in STATS19, which suggests a considerable degree of misclassification. This proportion has varied little over time (Figure 4.9), with changes broadly in line with the ratio of serious to slight injuries in the complete STATS19 dataset (Figure B4, Annex B).

[24] It is necessary to assume that the characteristics of the links made reflect those of the missed matches, at least in broad terms.

[25] It could also be that the linkage is incorrect in some cases which should be borne in mind, although any such incorrect linkages are unlikely to invalidate the analysis presented here.

Figure 4.9 Proportion of linked records coded serious in STATS19



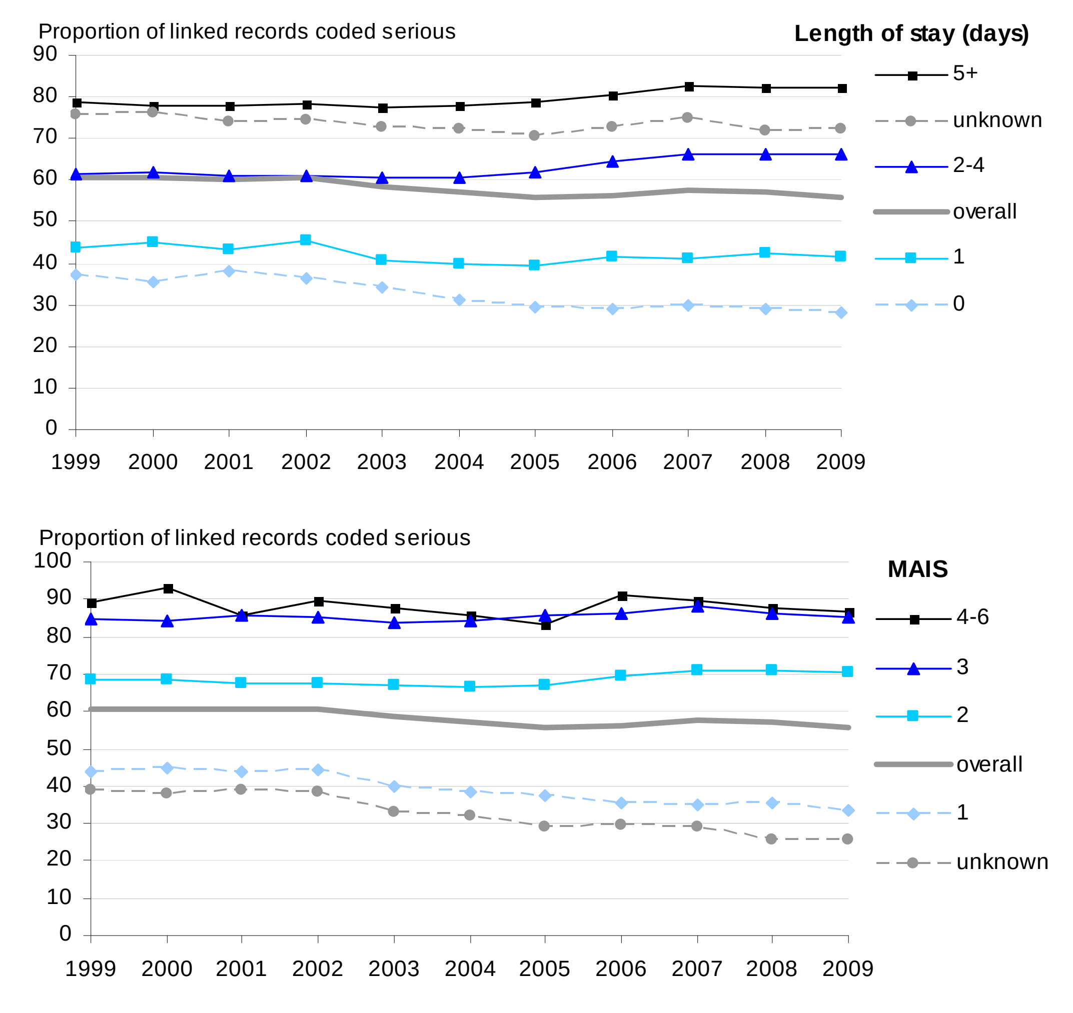
	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
—◆— % Serious	60	61	60	61	58	57	56	56	58	57	56

Note that this proportion can be influenced by hospital practices, and factors, such as availability of beds, may affect the propensity of a casualty to be admitted as well as the nature of their injuries.

In the linked dataset the police classification can be compared with alternative measures of severity derived from hospital data, including MAIS and length of stay (Figure 4.10). As might be expected, the proportion coded serious by police is higher for those with longer spells in hospital, or higher MAIS scores, being around 79% for those admitted for 5 or more days, and 86% for casualties with MAIS 3 or above. The most notable trends are a decline in the proportions coded serious for the least severely injured (length of stay 0 days or MAIS 1) which may indicate an increasing tendency for hospitals to admit casualties with less severe injuries. However, it is hard to conclude this with certainty.

This suggests that some of the linked casualties admitted and recorded as slightly injured by police may have relatively minor injuries, and only fall within the STATS19 serious definition by virtue of being admitted (e.g. for observation). So the degree of misclassification may be overstated.

Figure 4.10 Proportion of linked casualties coded as seriously injured in STATS19 data by i) length of stay in hospital and ii) MAIS level

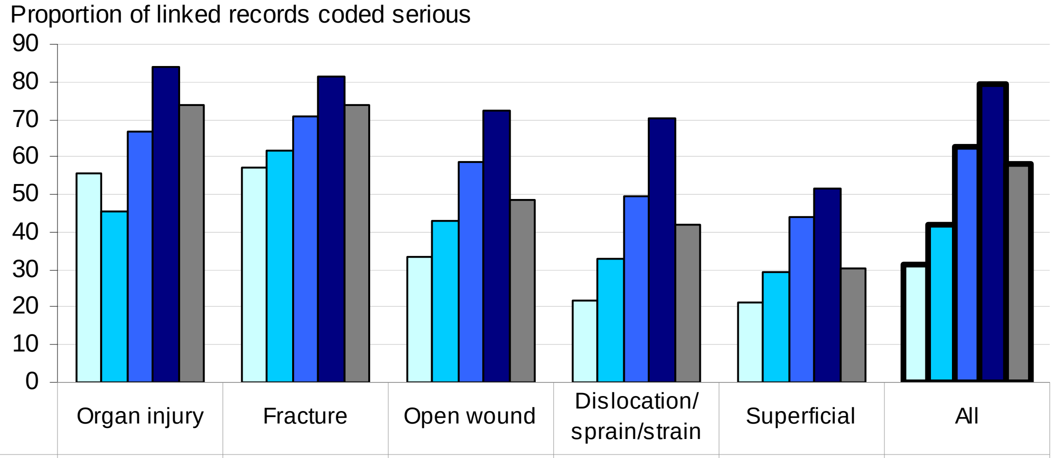


Overall, 6% of linked slight casualties have MAIS 3 or above with 12% admitted for 5 days or longer, compared with 27% and 34% of linked serious casualties (Table B7, Annex B).

Police recording of severity varies with primary diagnosis injury type found in HES (Figure 4.11). Even after allowing for the length of stay in hospital, a lower proportion of casualties admitted to hospital with a primary diagnosis of dislocations, sprains or strains are coded as seriously injured in comparison with fractures and internal injuries.

Dislocation/sprain/strains and superficial injuries to the neck are particularly likely to be misclassified as slight injuries by police (23% coded serious overall, Table B8), although this may reflect the fact that such injuries are less likely to be severe.

Figure 4.11 Proportion of linked casualties coded as seriously injured in STATS19 data by nature of injury recorded as primary diagnosis in HES, 1999-2009



	Organ injury	Fracture	Open wound	Dislocation/ sprain/strain	Superficial	All
□ 0 days	55	57	34	22	21	31
■ 1 day	46	62	43	33	29	42
■ 2-4 days	67	71	58	50	44	63
■ 5+ days	84	81	72	70	51	79
■ All	74	74	49	42	31	58

A higher proportion of motorcyclist and pedal cyclist casualties are (correctly) recorded as seriously injured in the linked dataset (Table 4.4). These groups are more likely to have more severe injuries, which in turn are more likely to be classified correctly by police, so this is not surprising.

Table 4.4 Proportion of linked casualties coded as seriously injured in STATS19 data by road user type, HES years 1999 - 2009

Road user type	Serious	Slight	Total	*Percent serious*
Car	43,606	39,725	83,331	*52*
Motorcycle	25,307	11,303	36,610	*69*
Pedal cycle	8,509	6,696	15,205	*56*
Pedestrian	29,662	19,023	48,685	*61*
All	**111,220**	**80,222**	**191,442**	***58***

The proportion of matched records coded as serious in STATS19 within each broad type of injury do not vary much for the main road user groups (Figure B5). The exception was less accurate coding of severity for dislocation injuries for car occupants.

4.3.2 Severity measures in HES: length of stay and MAIS

While the police data gave a broad indication of injury severity of casualties, the hospital data offers alternative indicators of severity for those casualties whose records have been matched to the hospital data.

Unsurprisingly, there is an association between MAIS and length of stay at hospital (Table 4.5). Overall, 66% of MAIS 1 cases spend 1 day or less in hospital; conversely 71% of MAIS 3 casualties are admitted for 2 or more days (with 22% having unknown length of stay).

Table 4.5 MAIS level of by length of stay for linked casualty data, 1999 - 2009

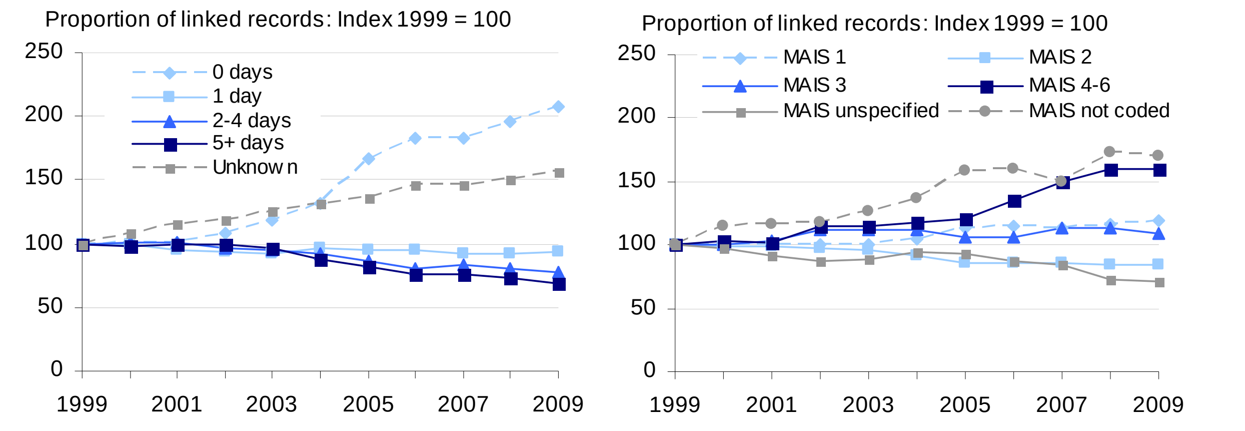
	Length of stay (days)					
MAIS	**0**	**1**	**2-4**	**5+**	**Unknown**	**All**
1	13,419	19,310	10,104	3,186	2,882	**48,901**
2	5,871	13,929	22,953	24,773	9,599	**77,125**
3	711	1,363	4,873	15,909	6,438	**29,294**
4-6	310	290	631	1,989	1,651	**4,871**
9 (Unspecified)	5,959	8,505	3,450	1,210	1,056	**20,180**
99 (Not coded)	4,039	3,558	1,750	931	793	**11,071**
All	**30,309**	**46,955**	**43,761**	**47,998**	**22,419**	**191,442**
Length of stay distribution for each MAIS level (row percentages)						
1	*27*	*39*	*21*	*7*	*6*	***100***
2	*8*	*18*	*30*	*32*	*12*	***100***
3	*2*	*5*	*17*	*54*	*22*	***100***
4-6	*6*	*6*	*13*	*41*	*34*	***100***
9 (Unspecified)	*30*	*42*	*17*	*6*	*5*	***100***
99 (Not coded)	*36*	*32*	*16*	*8*	*7*	***100***
All	***16***	***25***	***23***	***25***	***12***	***100***
MAIS distribution for each length of stay group (column percentages)						
1	*44*	*41*	*23*	*7*	*13*	***26***
2	*19*	*30*	*52*	*52*	*43*	***40***
3	*2*	*3*	*11*	*33*	*29*	***15***
4-6	*1*	*1*	*1*	*4*	*7*	***3***
9 (Unspecified)	*20*	*18*	*8*	*3*	*5*	***11***
99 (Not coded)	*13*	*8*	*4*	*2*	*4*	***6***
All	***100***	***100***	***100***	***100***	***100***	***100***

One possible advantage of using hospital-based measures of injury severity is in making comparisons of injuries across countries, and this has been considered within the EU for the SafetyNet project [11][26]. It was proposed that MAIS was a more reliable measure for monitoring trends in the more serious injuries compared to length of stay, as the latter is more susceptible to reflecting changes in hospital practices, for example an increasing tendency to admit patients, or reductions in treatment time for given injuries.

[26] However, even if an international classification such as MAIS were to be used, there would still be difficulties in making direct comparisons between countries. For example, some countries use the ICD-9 coding of injury diagnosis and there is some evidence (e.g. [11]) that use of ICD-10 coding results in generally lower MAIS scores, on average.

Figure 4.12 shows the trend in the *proportion* of linked records by both MAIS and length of stay[27] between 1999 and 2009. Although care is needed in interpreting the patterns shown, there is some evidence of increases in the proportions of inpatients with relatively more severe injuries (using MAIS). However, this may be partly due improvements in the recording of diagnosis and subsequently a decrease in the proportion of linked records with unspecified MAIS. Concurrently, the proportion of inpatients with stays of 0 days (i.e. not overnight) has more than doubled over this period.

Figure 4.12 Proportion of linked records by MAIS and length of stay: Index 1999 = 100

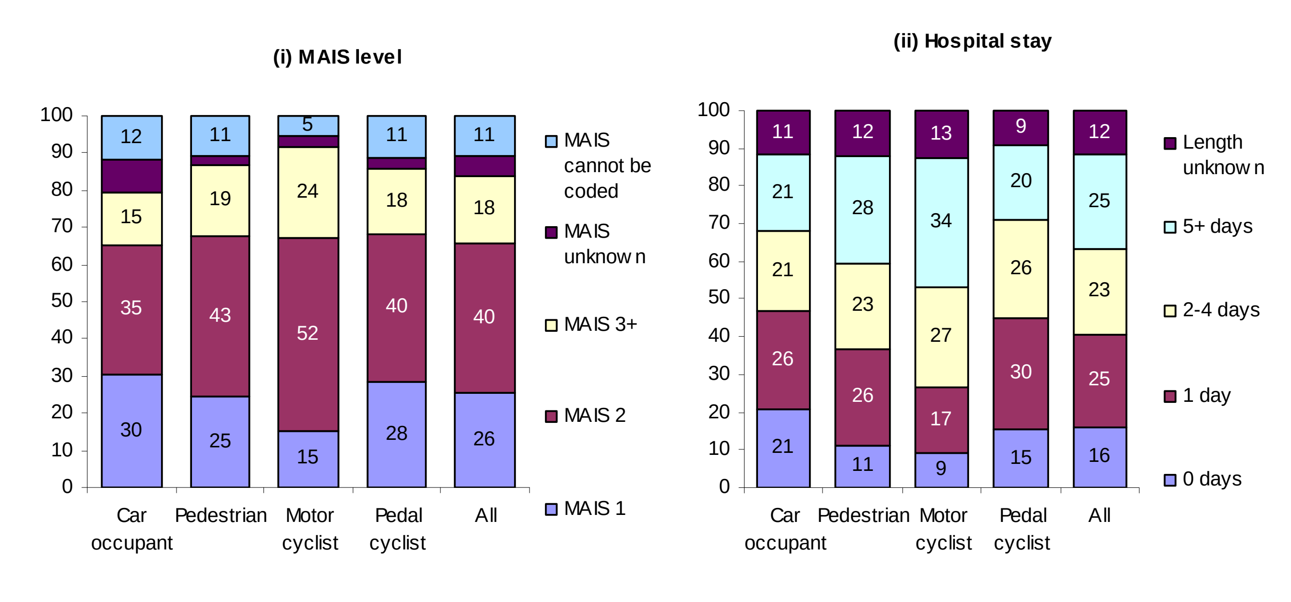


These trends may reflect a number of factors. There may also be genuine changes in distribution of severity of injury (for example, with some who would previously have died from their injuries may now survive but with severe injuries). However, other external factors such as increased use of short stay ‘observation’ wards[28] may result in an increasing tendency to admit for short periods, whilst improvements in treatment and other factors such as the availability of beds may reduce the time that those with more severe injuries spend in hospital.

Figure 4.13 shows the distribution of HES recorded casualty type by both MAIS level and length of stay in the linked dataset. The broad patterns by the two different proxy measures of severity shown are similar. Motorcyclists and pedestrians admitted to hospitals have a higher proportion of more severe injuries than other road user groups using either proxy measures of severity[29].

[27] Note that these are trends in proportions, rather than numbers, because the total number of records has not been adjusted for variations in quality of linkage which changes over time
[28] Sometimes known as Clinical Decision units
[29] It is worth noting that the unknown cases are likely to reflect different groups for the two measures. Cases where MAIS is not coded are those where there is no suitable injury diagnosis coded, and may therefore include those who have no injury – in general, such cases will probably be less severely injured, on average. Conversely, a missing length of stay in this dataset represents an inpatient having more than one episode of care in hospital. On average, such inpatients are likely to spend longer in hospital (and have more severe injuries) than those having a single episode of care.

Figure 4.13 Distribution of linked records by i) MAIS level and ii) length of stay in hospital, matched dataset 1999-2009



4.3.4 Summary and discussion

The linked STATS19-HES dataset contains around 190,000 records, covering road casualties admitted to hospitals in England and recorded in the police database for the period 1999 to 2009. The analysis here suggests that given that police officers are typically required to code severity of injury within a short time of the accident, without extensive medical knowledge, their classification of the injury severity is reasonably good:

- Overall, 58% of linked casualties are correctly recorded as seriously injured in the STATS19 dataset. However, the proportion coded serious is considerably higher for those spending longer in hospital (79% for those admitted for 5 days or more) and those having more severe injuries (86% for casualties with MAIS 3 or above). Some of those miscoded by police appear to have relatively minor injuries and only fall within STATS19 serious definition by virtue of being admitted to hospital (e.g. for observation). So the degree of misclassification may be overstated.
- There is no clear evidence of a systematic deterioration in the accuracy of coding of injury severity over the past decade, based on these results. Although the proportion of linked records coded serious falls marginally, this may reflect a fall in serious injuries relative to slight in the whole STATS19 dataset.
- The proportion of linked records correctly coded as serious by police is lower for dislocations, strains and sprains (42%) and for ‘superficial’ injuries (31%). The coding of casualties with a dislocation/sprain/strain to the neck is particularly difficult, with only 24% being recorded as seriously injured in STATS19.
- Measures of injury severity derived from the data available in HES (length of stay and MAIS) indicate that among the linked casualties motorcyclists are

likely to have on average the most severe injuries and spend longer in hospital.

These findings are generally not surprising or new, confirming the results of previous work. In terms of the coding of injury severity, the Simpson study [12] based on 1993 data suggested around 60% of inpatients were coded as seriously injured by police. Other studies [3] of individual hospitals have found broadly similar proportions of misclassification by police. In addition, there is a reasonable degree of correspondence between the results of the linkage carried out for England and Scotland [10], once the differences in the severity distribution trends are taken into account.

In terms of trends over time, particularly in the coding by police, the results presented above are difficult to interpret. Past studies (e.g. [19]) have reported a fall in the proportion of inpatients correctly coded as seriously injured. The proportion in the linked dataset for England does fall over the period studied, but it does not show a clear, steady decline. In addition, as discussed previously, some casualties with relatively minor injuries only fall within STATS19 serious definition by virtue of being admitted to hospital.

The above analysis has concentrated on measures of severity, but the linked dataset may also be useful in assessing the quality of recording of common variables, for example road user type and casualty class. Annex C contains some details of such analysis, which broadly suggests that there is a good degree of agreement between STATS19 and HES on these variables.

5. Estimating serious casualties using linkage results

5.1 Introduction

The previous section explored the characteristics of the linked dataset. This section presents an application of the results of linkage to estimate levels and trends of serious road casualties in England, and by scaling, Great Britain (GB).

Whilst most, if not all, road deaths become known to police, it has long been known that the STATS19 dataset provides an incomplete count of non-fatal casualties. This is clear from comparison with other datasets, and illustrated by the analysis presented in the preceding section.

In recent years, attempts have been made to estimate the true number of road casualties in Great Britain, largely based on survey data [8,24]. However, it is also possible to produce estimates based on the linkage results, using a technique known as the capture recapture method.

Police and hospital data have also shown different trends during some calendar periods. There are many possible reasons for this, and the differences have been explored in past reports (e.g. [2],[3],[25]). Although the results of this linkage cannot provide any definitive answers, they offer scope to add to our understanding.

5.2 Key assumptions

In order to carry out the analysis in this section, it is necessary to adjust the number of linkages achieved to allow for the likely variations in quality of linkage over time. In particular, the probable underestimation of the number of true matches which is likely to be decreasing over time (as the recording of STATS19 postcode improves). This adjustment is outlined below.

Throughout this section, the following assumptions are made:

- *Accuracy of linkage.* It is assumed that, after adjustments for variations in the post coding recording over time, the linkage is largely correct and that there are few missed or incorrectly linked records.
- *Suitability of adjustment*. In particular, it is assumed that the characteristics of the missed matches (which are allowed for by the scaling factors applied) are similar to those of the achieved linkages.
- *Accuracy of coding*. The following calculations use STATS19 serious records and those in HES that are classed as within scope of STATS19. In practice there will be other records within the file that are miscoded. These are allowed for to some extent in the calculations but the precise effect is uncertain.
- *Appropriateness of the capture-recapture method* when applied to estimating number of road traffic casualties (discussed below). The key point is that some of the key assumptions may not hold in this context so care is needed.

Clearly, the extent to which the following conclusions hold depends on the validity of the above assumptions. In general, there are doubts regarding each which cast some doubt on the results and mean that the conclusions should be treated as broad indications.

5.3 Adjustment of achieved linkages

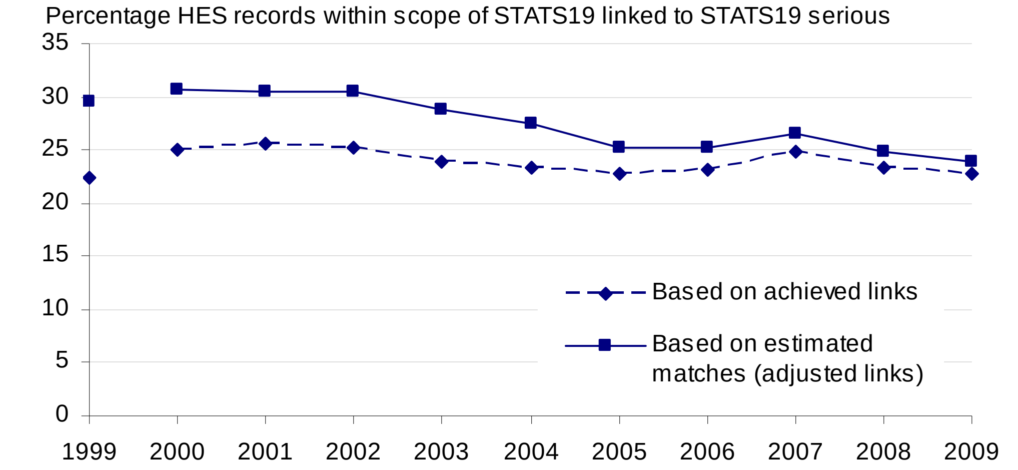
A very simple method of adjustment has been used here, which allows for the variation in availability of postcode information over time[30]. The adjusted number of links can be considered as the number we might have expected to achieve with full recording of postcode on STATS19, though it does not allow for other factors (which may vary over time) and therefore probably still represents an under-estimate of the true number of common records. Thus this scaling should improve the reliability of conclusions based on the following analysis, but it cannot be considered a fully robust approach and this should be borne in mind. The table sets out details of the adjustment, and the chart shows the trend in achieved and adjusted links as a proportion of the HES file (note this is based on only links to STATS19 serious records and accidents within scope of STATS19 in HES).

Table 5.1 Achieved and adjusted linked records as a proportion of STATS records (thousands)

Achieved links: HES in scope of S19 to S19 serious	8.6	9.5	9.6	9.3	9.3	9.3	9.6	9.5	10.1	9.2	9.3	**103.2**
Proportion S19 serious linked	*25*	*29*	*30*	*30*	*32*	*34*	*38*	*38*	*42*	*42*	*44*	***34***
Estimated proportion linked - postcoded records	*34*	*35*	*35*	*36*	*38*	*40*	*42*	*42*	*45*	*44*	*46*	***40***
Adjusted links - assuming all postcoded	11.4	11.6	11.4	11.3	11.1	10.9	10.6	10.4	10.8	9.8	9.8	**119.1**
Adjustment (%)	*32*	*23*	*19*	*21*	*20*	*17*	*11*	*9*	*7*	*6*	*6*	***19***

[30] An alternative approach might be to carry out a more detailed analysis of false positives and false negatives using the approach of Annex A and use this as a basis for adjustment. Results are similar but not identical.

Figure 5.1 Proportion of HES records linked to STATS19 serious



5.4 Estimating serious casualties

5.4.1 The capture recapture method

The capture-recapture method provides a means of estimating the (unknown) size of a population based on two or more counts, or registers. Typically the method has been used in zoology to provide counts of animal populations, but the same approach has been applied to other fields, including estimation of road traffic injuries. It is not the intention of this report to provide a detailed description of the method (see for example for further details [16], [17], [18]).

Table 5.2 illustrates the cross-tabulation of data from two sources, in this case police and hospital datasets. The total number of road casualties in the population is represented by n, the total number of road casualties reported to the police and recorded in STATS19 is represented by n(P) and the total number of road traffic casualties admitted to hospital and recorded in HES is presented by n(H). Road casualties who are not recorded in either HES or STATS19 are presented by n(not P, not H). The overlap of casualties that are in both of these sources, or casualties that appear in both STATS19 and HES are presented by n(P,H). The number of road casualties who are reported in STATS19, but who are not admitted to hospital are represented by n(P, not H). Similarly, the number of road casualties who are recorded in HES, but not reported in STATS19 is represented by n(not P, H).

Table 5.2: Cross-tabulation of data from police and hospital datasets

		Hospital		
		Yes	No	
Police	Yes	n(P,H)	n(P,notH)	n(P)
	No	n(notP, H)	n(notP,notH)	
		n(H)		n

Using the figures described in Table 5.2, the total number in the population (n) can be derived using the following formula:

Total number of the population = n = n(P)*n(H)/n(P,H)

To perform this calculation, it remains only to estimate n(P,H) – the overlap between the two sources, which can be obtained from the results of the linkage.

This approach relies on a number of assumptions. In summary these are:

1. *Closed population*. No entry or loss among road traffic casualties – for a study such as this one covering the whole of England this should broadly hold.

2. *Perfect identification of common records*. In cases where data from two sources is linked at record level, as here, this assumption basically requires perfect linkage, with no missed or false matches. This is clearly unrealistic, although the adjustment described attempts to address this to some extent.

3. *Independence of data sources.* A key assumption underlying the calculation of the capture-recapture estimate is that the two data sources are independent, so that, for example, the proportion of all serious casualties that are admitted to hospital is the same as the proportion admitted among those known to police. This is unlikely to be true. For example, in many cases casualties will be referred to hospital by police attending the scene of an accident, creating a positive relationship (i.e. those known to police are more likely to also be admitted) which leads to the resulting estimate being biased downwards.

4. *Homogeneity of capture*. This means that all casualties should have the same probability of becoming known to police (and of being admitted to hospital, although the probability of admission does not need to be the same as the probability of appearing in the police data). Previous work and the results presented in Section 4 demonstrate that this is not the case. For example, the probability of appearing in either dataset is likely to be related to severity of injury (within the serious category), with those more seriously injured, more likely to be captured. Variations, for example by road user and collision type, can be allowed for to some extent using stratification. This was done by calculating the capture-recapture estimate separately for sub-groups where the probability of inclusion is more likely to be similar[31]. This reduces, but does not eliminate, the degree to which the assumption does not hold.

5. *Same geographical area and time period covered by both sources*. In this case, this assumption can be considered to be practically met – although there will be some differences (for example, where hospital admission is not on the same day or in the same country as the accident) these are likely to affect only a relatively small proportion of cases.

6. *Perfect identification of subjects of interest*. The criteria for defining a subject of interest should be precise, and should be the same for both datasets. In the following, the population of interest is seriously injured casualties in road traffic

[31] It is possible to use more sophisticated statistical modelling to deal with the problem of heterogeneous inclusion probabilities; however, this approach will not be applied here. See [16] for more details.

accidents in England, based on the STATS19 definitions used by the police[32]. Note that this definition is wider than hospital admission, so that we would not expect the HES dataset to provide full coverage of serious casualties. The capture-recapture method cannot be used to estimate the total number of non-fatal injuries (including those with slight injuries) in England as there is no suitable hospital dataset covering less severe injuries[33].

In conclusion, several of the key assumptions of capture-recapture do not hold when applied to estimation of road traffic injuries using police and hospital datasets – in particular independence, homogeneity and perfect identification of common records. The degree of violation can be minimised to some extent, for example, by calculating the capture-recapture estimates separately for different sub-groups, within which casualties are more homogenous. Despite this, the estimates presented here should therefore be considered as broadly illustrative, rather than precise figures.

5.4.2 Application to linked data

Table 5.3 presents the capture-recapture estimates for the average number of serious casualties in England over the period 1999-2009. These figures were based on the adjusted linkage results. To allow for the heterogeneity of inclusion, separate estimates were calculated for subgroups defined using road user type, nature of collision and age group.

These figures are based on all accidents within HES classed as within scope of STATS19. This includes a number of groups of casualties that in practice are less likely to become known to police, for example pedal cycle casualties in accidents involving no motor vehicle.

For the period 1999-2009, STATS19 records an average of 27 thousand serious casualties in England (and 32 thousand in GB). By comparison:

- Overall, the estimated average number of serious casualties in England over the period 1999 to 2009 is 91 thousand. By a simple scaling, this would suggest nearly 104 thousand casualties in Great Britain[34].
- Excluding non-collision pedal cycle casualties, the estimated number of serious casualties is broadly 83 thousand in England (96 thousand in GB).
- Restricting to only traffic accidents as defined in HES, which excludes for example those recorded as ‘boarding and alighting’[35], results in an estimate of 78 thousand (90 thousand for GB).

[32] The definition of serious injury is that given in section 4.3

[33] Data on attendances at Accident & Emergency in England is currently published as experimental statistics, and may allow such estimation in future. However, unlike the inpatient data, it is not possible to easily identify road traffic casualties within this dataset.

[34] Note that these figures should be treated as broad illustrations; besides the assumptions discussed above, some cases with e.g. unknown age in STATS19 or unknown road user type in HES were excluded and so the estimates can not be considered precise.

[35] Non-collision pedal cycle accidents are also omitted though they fall within the definition of traffic accidents when on the public highway. Figures are based on an equivalent calculation but the details are omitted here.

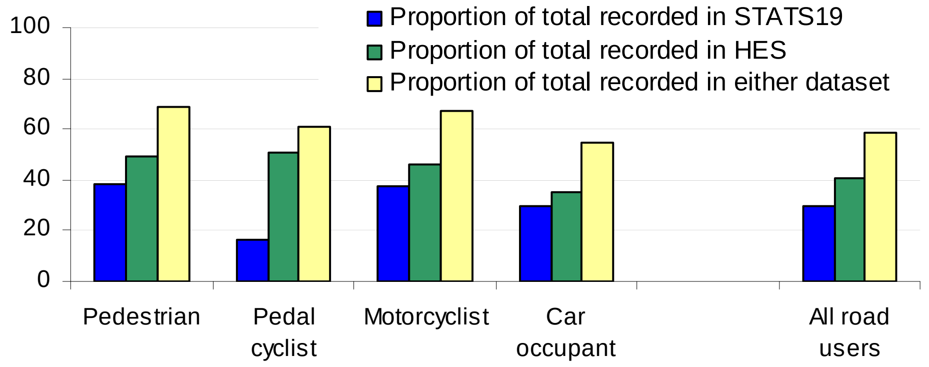
Table 5.3: Capture-recapture estimate of serious casualties in England and Great Britain, 1999-2009 avg.[36]

			STATS19 Serious	HES (within scope of STATS19)	Linked records (after adj.)	Estimated total (England)	% total in STATS19	% total in HES	% total in either	Estimated total (Great Britain)
Pedestrians		**Total**	**6,000**	**7,730**	**2,990**	**15,700**	***38***	***49***	***68***	**18,000**
		0-15	2,010	2,720	1,130	4,800	*42*	*57*	*75*	5,500
		16-59	2,850	3,470	1,210	8,100	*35*	*43*	*63*	9,300
		60+	1,130	1,540	640	2,700	*42*	*57*	*75*	3,100
Pedal cyclists		**Total**	**2,200**	**6,750**	**810**	**13,300**	***17***	***51***	***61***	**15,200**
	Motor vehicle involved	*Sub-total*	*2,030*	*1,910*	*720*	*5,400*	*38*	*35*	*60*	*6,200*
		0-15	490	560	220	1,200	*41*	*47*	*69*	1,400
		16-59	1,370	1,170	420	3,800	*36*	*31*	*56*	4,400
		60+	170	190	80	400	*43*	*48*	*70*	500
	No motor vehicle	*Sub-total*	*160*	*4,840*	*90*	*7,800*	*2*	*62*	*63*	*9,000*
		0-15	20	2,260	20	2,500	*1*	*90*	*90*	2,900
		16-59	130	2,150	60	4,600	*3*	*47*	*48*	5,300
		60+	20	430	10	700	*3*	*61*	*63*	800
Motorcyclists		**Total**	**5,520**	**6,830**	**2,420**	**14,800**	***37***	***46***	***67***	**17,000**
	Multi vehicle accident	*Sub-total*	*4,200*	*3,630*	*1,710*	*8,900*	*47*	*41*	*69*	*10,200*
		0-15	60	90	30	200	*30*	*45*	*60*	300
		16-59	4,010	3,390	1,620	8,400	*48*	*40*	*69*	9,600
		60+	130	140	60	300	*43*	*47*	*70*	300
	Single vehicle accident	*Sub-total*	*1,320*	*3,210*	*720*	*5,900*	*22*	*54*	*65*	*6,800*
		0-15	20	190	10	400	*5*	*48*	*50*	400
		16-59	1,260	2,870	680	5,300	*24*	*54*	*65*	6,100
		60+	40	150	20	200	*20*	*75*	*85*	300
Car occupants		**Total**	**11,990**	**14,090**	**4,180**	**40,100**	***30***	***35***	***55***	**46,000**
	Multi vehicle accident	*Sub-total*	*8,470*	*8,980*	*2,790*	*27,400*	*31*	*33*	*54*	*31,400*
		0-15	420	590	130	1,900	*22*	*31*	*46*	2,200
		16-59	6,620	6,630	2,070	21,200	*31*	*31*	*53*	24,300
		60+	1,430	1,760	580	4,300	*33*	*41*	*61*	5,000
	Single vehicle accident	*Sub-total*	*3,520*	*5,100*	*1,390*	*12,700*	*28*	*40*	*57*	*14,600*
		0-15	150	260	50	800	*19*	*33*	*45*	900
		16-59	3,090	3,900	1,200	10,100	*31*	*39*	*57*	11,500
		60+	280	940	140	1,900	*15*	*49*	*57*	2,200
Others		**Total**	**1,360**	**1,810**	**350**	**7,200**	***19***	***25***	***39***	**8,200**
		0-15	70	100	20	500	*14*	*20*	*30*	500
		16-59	1,010	1,100	260	4,200	*24*	*26*	*44*	4,900
		60+	280	610	70	2,500	*11*	*24*	*33*	2,900
All road users		**Total**	**27,060**	**37,210**	**10,750**	**91,100**	***30***	***41***	***59***	**104,400**
		0-15	3,260	6,770	1,610	12,300	*27*	*55*	*68*	14,100
		16-59	20,330	24,690	7,530	65,700	*31*	*38*	*57*	75,300
		60+	3,480	5,760	1,610	13,100	*27*	*44*	*58*	15,000
All excluding pedal cyclist accidents with no other vehicle		**Total**	**26,900**	**32,370**	**10,660**	**83,200**	***32***	***39***	***58***	**95,500**
		0-15	3,240	4,510	1,590	9,800	*33*	*46*	*63*	11,200
		16-59	20,210	22,530	7,470	61,100	*33*	*37*	*58*	70,100
		60+	3,460	5,330	1,600	12,300	*28*	*43*	*58*	14,200

[36] Note those records with missing age (in both files) and missing accident type (in HES) are excluded from the figures used for both linked and unlinked records. Cases where collision type is unknown (in HES) are apportioned pro-rata to those where it is recorded.

The estimated proportions of total casualties recorded in STATS19, in HES and in either STATS19 or HES are shown in Figure 5.2 for the main road user groups.

Figure 5.2 Estimated proportion of total serious casualties known to police or admitted to hospital, based on capture-recapture estimate for England 1999-2009



- Overall, these estimates suggest that around a third of total serious casualties become known to the police. There are higher reporting rates for pedestrians and lower rates for cyclists (largely explained by accidents with no motor vehicle).
- Police become aware of a higher proportion of motorcyclist and car occupant casualties occurring in multi-vehicle accidents, compared with single vehicle accidents (although the difference is relatively small for car occupants).
- For pedestrian and pedal cyclist casualties, a higher proportion of children and those aged 60 or over are found in STATS19. The reverse is true over all road user groups.
- Around 40 per cent of the estimated total serious casualties are found in HES i.e. are admitted to hospital – including around half of pedestrian and pedal cyclist casualties. These estimates suggest that around 40 per cent of total serious casualties may not appear in either dataset

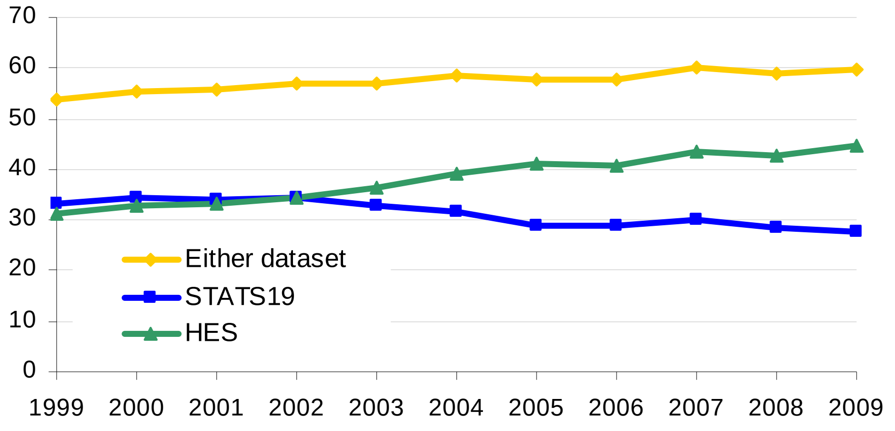
5.5 Estimation of trends

It is possible to apply the capture-recapture method to produce estimates for single years and thus estimate trends. As the sample size for some subgroups becomes very small, only the overall road user group totals are presented in Table 5.4. Given the limitations of the method, these figures should be considered very broadly illustrative rather than precise. Figure 5.3 shows how the estimated proportion of total serious casualties in England captured in the STATS19 and HES data changes over the period studied.

Table 5.4: Estimated serious casualties based on capture-recapture, England 1999 –2009 (thousands)

	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Pedestrians	**17.9**	**16.6**	**17.1**	**16.2**	**15.5**	**15.4**	**16.2**	**16.2**	**15.8**	**15.2**	**14.3**
% STATS19	*42*	*44*	*41*	*41*	*39*	*37*	*34*	*33*	*34*	*33*	*33*
% HES	*46*	*48*	*47*	*46*	*48*	*50*	*49*	*47*	*49*	*49*	*52*
% Either	*68*	*71*	*68*	*68*	*68*	*68*	*66*	*64*	*66*	*66*	*67*
Pedal cyclists	**6.1**	**5.8**	**5.3**	**4.9**	**5.2**	**5.3**	**5.6**	**5.8**	**5.7**	**6.2**	**6.0**
% STATS19	*40*	*37*	*40*	*39*	*37*	*34*	*33*	*33*	*35*	*33*	*36*
% HES	*33*	*31*	*33*	*32*	*32*	*33*	*36*	*34*	*36*	*34*	*39*
% Either	*60*	*57*	*59*	*59*	*57*	*55*	*57*	*56*	*59*	*56*	*61*
Motorcyclists	**14.6**	**15.5**	**15.7**	**16.1**	**16.6**	**15.4**	**16.0**	**15.4**	**14.6**	**14.1**	**14.2**
% STATS19	*38*	*38*	*38*	*38*	*38*	*35*	*33*	*34*	*37*	*34*	*33*
% HES	*39*	*40*	*40*	*41*	*44*	*45*	*47*	*47*	*50*	*49*	*51*
% Either	*63*	*64*	*63*	*64*	*66*	*65*	*65*	*66*	*69*	*67*	*68*
Car occupants	**50.3**	**46.0**	**45.2**	**44.0**	**42.2**	**40.0**	**39.6**	**39.0**	**34.8**	**33.7**	**33.3**
% STATS19	*31*	*32*	*32*	*32*	*30*	*30*	*27*	*27*	*28*	*26*	*25*
% HES	*26*	*28*	*29*	*30*	*33*	*36*	*39*	*39*	*42*	*42*	*44*
% Either	*49*	*52*	*52*	*53*	*53*	*55*	*56*	*56*	*58*	*57*	*58*
All road users	**98.2**	**92.5**	**91.4**	**88.7**	**86.4**	**82.4**	**84.6**	**83.4**	**77.8**	**75.8**	**74.3**
% STATS19	*33*	*35*	*34*	*34*	*33*	*32*	*29*	*29*	*30*	*29*	*28*
% HES	*31*	*33*	*33*	*35*	*37*	*39*	*41*	*41*	*44*	*43*	*45*
% Either	*54*	*56*	*56*	*57*	*57*	*59*	*58*	*58*	*60*	*59*	*60*

Figure 5.3 : Proportion of estimated total recorded in STATS19 and HES 1999 – 2009

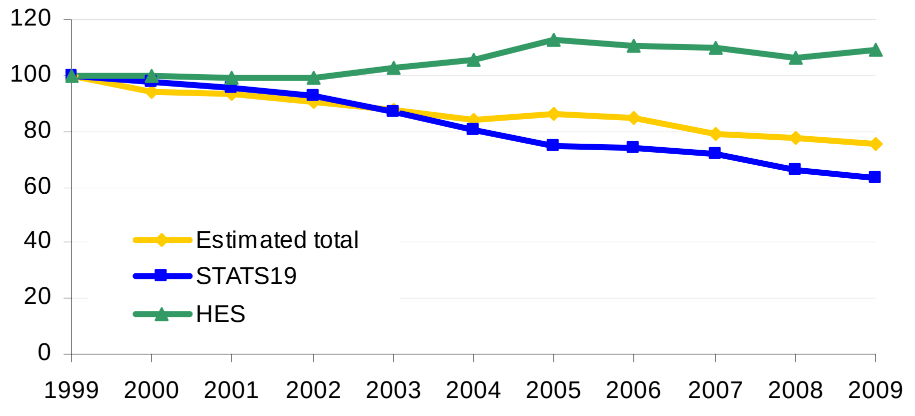


The estimated proportion of the total known to police is relatively consistent, falling between 2002 and 2005. This could be the result of an increasing tendency among hospitals to admit casualties with less severe injuries (that may previously not have been admitted)[37] rather than a change in police recording practice over this period. Thus, we can tentatively conclude that the proportions of total casualties known to police have very broadly stay constant over the period 1999 to 2009. That is, given the assumption of hospital practices changes- there is no clear evidence of a change in the level of reporting to police, at least at the national level.

Figure 5.4 compares the trend in serious casualties shown by STATS19 and HES road casualty admissions within scope of STATS19 with the capture-recapture estimate. Bearing in mind the limitations discussed above, this suggests that the trend shown by STATS19 is more likely to provide a better reflection of the trend in overall serious casualties than that shown by HES. However, again the limitations of the calculation prevent this from being anything more than a tentative conclusion.

Whilst certainly not conclusive, these estimates provide some support for the view that STATS19 provides the best single source of information on road casualty trends for England, at the national level. But understanding differences in the trends is not straightforward and it is not possible to draw firm conclusions on the basis of the available evidence. There are many possible reasons for differences, which are difficult to unpick. These are discussed further in, for example, [2], [24] and [25].

Figure 5.4 : Trend in serious casualties – estimated, STATS19 and HES (1999 = 100)



[37] See for example [2] and [24] for details of some of the factors affecting the HES data over this period

5.6 Illustrative example

There were an estimated 74.3 thousand seriously injured road casualties in England for 2009, excluding pedal cycle casualties in accidents involving no motor vehicle (Table 5.4). This would suggest there were around 85 thousand road casualties in Great Britain in 2009, excluding pedal cycle casualties in accidents involving no motor vehicle (assuming the casualty rate per population in England applies to the whole of Great Britain). Using this estimate, and information on the number of STATS19 records, the number of road traffic hospital admission and the number of records which were matched between STATS19 and HES, we can give a broad illustrative breakdown of the estimates number of seriously injured casualties by different sources as shown in Table 5.5. These figures exclude pedal cycle casualties in accidents involving no motor vehicle, but include the remaining types of accident which are within scope of the STATS19 definition.

Table 5.5: Illustrative estimate of the number of serious casualties in Great Britain in 2009

		HES road traffic		
		Yes	No	
STATS19 serious	Yes	a = 11000	b = 13000	f = 24000
	No	c = 28000	d = 34000	
		e = 39000		g = 85000

Figures exclude pedal cyclist casualties in accidents with no motor vehicle

In Table 5.5, **cell a** represents those casualties known to police and attending hospital as road traffic admissions, n(P,H). This relies on the accuracy of record linkage. If true matches are missed then they will not be included here but will instead appear in cells b and c (resulting in overestimation).

In Table 5.5, **cell b**, represents STATS19 serious casualties not admitted and not coded as road traffic casualties in HES, n(P, notH). This suggests that less than half of STATS19 serious casualties are admitted to hospital (around 45 per cent). However, this probably represents an underestimate as this cell may include cases where a casualty is recorded as serious in STATS19 and admitted to hospital, but miscoded in HES (e.g. as a non traffic accident[38], or with missing cause code).

In Table 5.5, **cell c**, represents casualties admitted to hospital following a road traffic accident within scope of STATS19, but not present in the STATS19 serious category, n(notP, H). As illustrated in Section 4, some of these casualties will appear in STATS19 but misclassified as slightly injured (very crudely around 8 of the estimated 28 thousand).

[38] The data linkage suggests that there are of the order of a thousand cases recorded as non-traffic accidents in HES that are linked to STATS19 serious records per year. Note that it is also possible that some of those coded as traffic accidents in HES (and not linked to STATS19) are misclassified, so perhaps the most reasonable assumption is that overall this misclassification cancels out – though there is little evidence on which to base this.

In Table 5.5, **cell d,** representing road casualties not recorded either as road traffic admissions in HES or STATS19 serious, n(notP, notH). However, this cell may include records miscoded by police as slightly injured. Crudely estimating by scaling up those in cell c, suggests around 10 thousand such casualties. In practice, this may be an underestimate as it is likely that the police classification will be more accurate for those relatively more seriously injured. This cell will also include some casualties admitted to hospital but not coded as traffic accidents.

Based on these estimates, very broadly just over a quarter (28 per cent) of the estimated total serious casualties becomes known to police (including those where severity of injury is miscoded). Note that there will also be some genuinely slightly injured casualties wrongly included among the serious total by police[39]. No adjustment has been made for this group here, which may result in an upward bias in the estimated total number of serious casualties.

5.6 Discussion and conclusion

If a reasonably reliable linkage can be established between the police and hospital data, calculating estimates of the total number of serious casualties in England (and by scaling, Great Britain) using the capture-recapture method is computationally straightforward.

Some of the key assumptions of capture-recapture do not hold when the method is used to estimate road traffic injuries. The overall effect is hard to assess, although the literature suggests that generally estimates will be undercounts (e.g. [16]). Despite the limitations this approach, it can be valuable both in illustrating the incompleteness of police data, and estimating trends which take account of data from both police and hospital sources, at least in broad terms.

While, there are some considerable limitations, the results using capture-recapture suggest that:

- Overall around a third of the estimated total serious casualties become known to police and recorded in STATS19 as serious, with around 40% admitted to hospital and included in HES as road traffic accidents.
- These estimates provide no strong evidence to suggest that the proportion of serious casualties known to police has changed over the last decade. It may have fallen slightly over the period studied. However, this may be due the steady increase in the proportion admitted, which could be the result of an increasing tendency for hospitals to admit casualties with relatively minor injuries.
- The trend over time in the estimated total serious casualties is more similar to the trend shown by police rather than hospital data, which supports the general conclusion that STATS19, though incomplete, is a more reliable source of data on trends in serious casualties.

Differences in coverage, types of dataset used and linkage methods make direct comparisons with other work difficult. Comparing to a study with a similar linkage approach in Scotland [10] suggests broadly similar estimates of serious casualties for Great Britain. However, the Scottish study suggests that a higher proportion of casualties are included in either the police or the hospital datasets.

[39] By definition, these would not be admitted so would appear in the second cell

A study of 15-24 year old serious casualties for Scotland estimated that police and hospital datasets were respectively two-thirds and three-quarters complete [19] –higher than the overall figures presented above. Similarly, application of the approach in a French study [16] estimated maximum ascertainment rates of 57 per cent for police data and 87 per cent for medical registry.

Simpson's study [12] from the early 1990s found that around half of serious road casualties attending hospital were admitted[40] and estimated that to take account of police and hospital recorded data the number of seriously injured casualties in national casualty data should be increased by a factor of 2.76 (equivalently, the police recorded figure represents around 36 per cent of the total known to police or hospital). This work informed the Department's best estimate of around 80 thousand serious injuries per year [8], and the estimates presented here are broadly comparable with this, although the estimates Some of the key assumptions of capture-recapture do not hold when the method is used to estimate road traffic injuries. The overall effect is hard to assess, although the literature suggests that generally estimates will be undercounts (e.g. [16]). Despite the limitations this approach, it can be valuable both in illustrating the incompleteness of police data, and estimating trends which take account of data from both police and hospital sources, at least in broad terms.

[40] Broadly in line with the estimate of around 40 per cent in this study

6. Discussion

This report has aimed to illustrate that linking STATS19 and HES data for England is feasible, and useful in illustrating the different strengths and limitations of the two datasets.

The majority of the findings presented here are not new, or particularly surprising. This work adds to the evidence base and is useful in providing confirmation that results of more localised studies appear to hold at national level.

The results of this analysis suggest patterns reported in less recent studies continue to hold in broad terms. In particular, the findings generally support the conclusions of the Department's most recent study of under-reporting of road casualties published in 2006 [3]. In addition, it supports the conclusions of work that compared different sources of data on road safety, including first attempts to estimate total casualties [8]. They are also very broadly in line with findings of the linkage of police and inpatient data for Scotland which was used to represent Great Britain in the EU SafetyNet project.

This analysis illustrates some of the difficulties in producing a definitive figure for the number of serious casualties, which is sensitive both to the definition used and assumptions made in producing estimates. As noted in previous studies [12], the scope for improving overall levels of reporting of accidents to police may be limited. So understanding the limitations and completeness of the police data and allowing for these when using the data is more realistic. The analysis of the linked data can help in this understanding, but without allowing any firm conclusions.

There are a number of broad areas where further work could be carried out, of which the first is likely to be the most important.

- The value in linking police and hospital datasets is the potential to provide information relating to accident circumstances and vehicles involved with medical consequences. Analysis of hospital data in particular requires considerable skill, so a wider exploration of the linked dataset by researchers to assess its potential to add value to the existing road safety evidence base would be particularly useful. A more detailed multivariable analysis could also be considered.
- Analysis of the hospital data has shown how it offers scope to provide more detailed information about severity of injury (for example using MAIS score) than is available in the police data. If possible, linkage should be carried out on an annual basis, to provide reasonably up to date information on trends in road casualties by severity for monitoring improvements in road safety and monitor any potential changes in the completeness or accuracy of the police data.
- Although a sufficiently robust approach has been developed for the purpose of the analysis presented in this report, further development of the linkage methodology is possible. Particular areas for exploration might include developing a consistent method across all countries in Great Britain, and extending to wider hospital casualties – including those attending A&E – should this become possible, in order to get a more complete picture than can be obtained by looking at inpatients alone.

Acknowledgements

The Department for Transport would like to thank the Office for National Statistics for help in developing the initial method to link the datasets, and the NHS Information Centre for providing extracts of the HES dataset and carrying out the linkage.

Data supplied by

The central, authoritative source of
health and social care information

NHS
The
Information
Centre
for health and social care

Glossary of terms and abbreviations

AIS	Abbreviated Injury Scale
DfT	Department for Transport
HES	Hospital Episode Statistics
ICD-10	International Statistical Classification of Diseases and Related Health Problems 10th Revision
ICD-10: non-traffic accidents	Any vehicle accident that occurs entirely in any place other than a public highway.
ICD-10: S and T codes	Injury, poisoning and certain other consequences of external causes
ICD-10: traffic accidents	Any vehicle accident occurring on the public highway [i.e. originating on, terminating on, or involving a vehicle partially on the highway]. A vehicle accident is assumed to have occurred on the public highway unless another place is specified, except in the case of accidents involving only off-road motor vehicles, which are classified as non-traffic accidents unless the contrary is stated.
ICD-10: V codes	External causes of morbidity and mortality - transport accidents
MAIS	Maximum Abbreviated Injury Scale
NHS	National Health Service
STATS19	Department for Transport national road accident database
STATS19: road traffic accidents	Road accidents involving personal injury occurring on the public highway (including footways) in which at least one road vehicle or a vehicle in collision with a pedestrian is involved and which becomes known to the police within 30 days of its occurrence. One accident may give rise to several casualties. “Damage-only” accidents are not included.
STATS19: seriously injured casualties	An injury for which a person is detained in hospital as an “in-patient”, or any of the following injuries whether or not they are detained in hospital: fractures, concussion, internal injuries, crushings, burns (excluding friction burns), severe cuts, severe general shock requiring medical treatment and injuries causing death 30 or more days after the accident.
STATS19: slightly injured casualties	An injury of a minor character such as a sprain (including neck whiplash injury), bruise or cut which are not judged to be severe, or slight shock requiring roadside attention. This definition includes injuries not requiring medical treatment.

References

[1] Road accident casualties: a comparison of STATS19 data with Hospital Episode Statistics, Department for Transport (2006).

[2] The use of hospital data on road accidents in Road Casualties Great Britain: 2006 Annual Report, Department for Transport (2007).

[3] Road Safety Research Report No. 69: Under-reporting of Road casualties Phase 1, Department for Transport (2006).

[4] Reporting of road traffic accidents in London: matching police STATS19 with hospital accident and emergency department data. Ward H., Robertson S., Townley K., Pedler A. Transport Research Laboratory (2007). http://www.tfl.gov.uk/assets/downloads/ReportingLevelsMatchingStats19andHospitalDataFullReport.pdf

[5] Linkage of STATS19 and Scottish Hospital In-patients Data – Analysis. for 1980–1995. Keigan, M., Broughton, J. and Tunbridge, R. J. Transport Research Laboratory (1999).

[6] National Statistics Methodological Series No.25: Methods for Automatic Record Matching and Linking and their use in National statistics, Gill L., Office for National Statistics (2001).

[7] Road Casualties Great Britain: 2007 Annual Report, Department for Transport (2008).

[8] Reported Road Casualties Great Britain: 2008 Annual Report, Department for Transport (2009).

[9] European Center for Injury Prevention,University of Navarra, Algorithm to transform ICD-10 codes into AIS 90 (98 update)

[10] Linkage of SHIPS and STATS19 data for Scotland. Unpublished report for SafetyNet project

[11] D.1.15. Estimation real number of road accident casualties Final Report on Task 1.5, SafetyNet project

[12] Simpson, H F (1996). Comparison of hospital and police casualty data: a national study. TRL Report 173

[13] STATS20: Instructions for the Completion of Road Accident Reports http://www.dft.gov.uk/collisionreporting/Stats/stats20.pdf

[14] Reuring M and N. Bos (2009). Seriously injured road crash casualties in the Netherlands in the period 1993-2008; the real number on in-patients with a minimum MAIS of 2. SWOV report. http://www.swov.nl/rapport/r-2009-12.pdf.

[15] ICD codes: http://apps.who.int/classifications/apps/icd/icd10online/ and DfT lookup table: http://webarchive.nationalarchives.gov.uk/20110503151558/http://www.dft.gov.uk/excel/173025/221412/221549/227755/503336/RCGB2009Article6.xls

[16] Estimating non-fatal road casualties in a large French county, using the capture-recapture method. Amoros E, Martin JL, Laumon B. Accid Anal Prev. 2007 May; 39(3):483-90.

[17] Capture-recapture: a useful methodological tool for counting traffic related injuries? Morrison A and Stone D, Injury Prevention, 2000.

[18] Children are not goldfish – mark/recapture techniques and their application to injury data. Jarvis S et al, Injury Prevention, 2000.

[19] An evaluation of police reporting of road casualties. S Jeffrey, D H Stone, A Blamey, D Clark, C Cooper, K Dickson, M Mackenzie, K Major *Injury Prevention* 2009;**15**:13-18

[20] Using Multiple Datasets to Understand Trends in Serious Road Traffic Casualties. Ronan Lyons, Heather Ward, Huw Brunt, Steven Macey, Roselle Thoreau, Maralyn Woodford. Accident, Analysis & Prevention 2008; 40: 1406-1410

[21] Changes in safety on England's roads: analysis of hospital statistics.
BMJ 2006;333:73 (8 July). (http://www.bmj.com/cgi/content/full/333/7558/73)

[22] HES FOR PHYSICIANS: A guide to the use of information derived from Hospital Episode Statistics. Royal College of Physicians, London and Unit of Health-Care Epidemiology, University of Oxford (March 2006)
http://www.uhce.ox.ac.uk/hessepho/reports/HESForPhysicians.pdf

[23] Health administrative data: Exploring the potential for academic research, St Andrews: Administrative Data Liaison Service. Garratt, E., Barnes, H. and Dibben, C. (2010)

[24] Reported Road Casualties Great Britain: 2008 Annual Report, Department for Transport (2010).

[25] Proposals to improve the reporting of road casualties. UK Statistics Authority Monitoring Brief 1/2011.

[26] UK Statistics Authority Assessment Report 4: Road Casualty Statistics (July 2009).

Annex A: Quality of linkage

This annex contains further details of the linkage quality assessment based on a sample of the full dataset.

False positives

To assess the number of false positives, an empirical method was used using 2004 data. This involved inflating the 2004 STATS19 file with all records from 2003 and 2005, with the years changed to 2004. Comparing the proportion of hospital records linked to the data from these other years, which are clearly false linkages, we can estimate the likely proportion of false positives. Results are shown in Table A1[41].

Table A1 Estimated false positive rates based on 2004 data

		Links with inflated STATS19 file (2003,04,05 records all labelled 2004)			
		Link to 2004 record			
Level	**Links (2004)**	Same HES record	Different HES record	Linked to 03/05 record	**Estimated false positive rate**
1	5,248	5,243	0	0	*0.0*
2	3,895	3,893	4	2	*0.1*
3	568	567	0	0	*0.0*
4	346	346	0	0	*0.0*
5	448	447	0	2	*0.4*
6	557	554	2	2	*0.4*
7	650	648	1	2	*0.3*
8	536	535	0	10	*1.8*
9	46	46	0	0	*0.0*
10	80	80	0	6	*7.0*
11	68	68	0	5	*6.8*
12	83	82	0	36	*30.5*
13	396	396	0	2	*0.5*
14	251	249	0	13	*5.0*
15	25	25	0	1	*3.8*
16	62	59	0	2	*3.3*
17	50	49	0	4	*7.5*
18	62	60	1	56	*47.9*
19	2,568	2,556	0	70	*2.7*
20	951	935	0	69	*6.9*
21	217	215	0	16	*6.9*
22	100	98	0	13	*11.7*
23	267	261	0	102	*28.1*
24	143	137	0	101	*42.4*
99 (rejected)	2,208	249	40	568	*66.3*
Total	**19,825**	**17,798**	**48**	**1,082**	***5.7***
Total exc. rejected	**17,617**	**17,549**	**8**	**514**	***2.8***

[41] Note that these results relate to a slightly different version of the linkage method than that finally adopted; this means that the 2004 figures do not agree exactly with those presented in the main report. However, the difference is small and unlikely to invalidate the broad conclusions here.

This suggests that overall the proportion of links that are false positives is likely to be around 3%, with some variation by agreement level. In particular, where there is exact agreement on postcode the false positive rate is less than 1%. For agreement levels with greater tolerances the false positive rate is greater (in particular at levels 12 and 18 where up to two days difference in date of accident and date of admission is allowed). As the number of linkages made at the less precise levels is relatively small, this does not appear to be of particular concern.

This assessment is based on one year's data with small numbers in some categories, and might be considered as an under-estimation of the number of false positives because it does not allow for the fact that casualties in the same accident are potentially more likely to be incorrectly linked (to another casualty in the same accident). However, given that around three quarters of road accidents involve a single casualty, this is unlikely to have a great effect.

False negatives

A broad assessment of the number of missed matches (or false non-matches) was made based on a probabilistic calculation, essentially a simplified version of the approach used in French studies [13]. In summary the approach is:

P(false negative)
= P(records not linked | true match)
= P(non-agreement on matching variables | true match)
= 1 – P(agreement on matching variables | true match)

Thus this approach involves estimating the probability of agreement on matching variables (within allowed tolerances) where a link has been made. We assume that agreement on one variable is independent of agreement on others so that the probabilities can be computed for individual variables and multiplied. However, this assumption is unlikely to be true for all cases since if the police are required to estimate one variable then they are more likely to estimate others.

The linking process contains essentially two sets of nested levels (those where a postcode is present on the police record and those where postcode is missing), to avoid double counting of missed matches the calculation is only made for these two distinct groups as a whole, and not for each individual agreement level as was done for false positives above.

Estimates of agreement probability within the allowed tolerances were made empirically, using 2004 data. For example, to calculate the agreement on age, data were linked using other variables not including age (postcode, gender and year in particular). The estimated probability of agreement was calculated as the proportion of cases where the casualty age in STATS19 and HES agreed for the resulting links. For this example, in 97.2% cases the difference in ages recorded in STATS19 and HES was no more than 3 years, where STATS19 postcode was available.

Similar estimates for other variables were made and these are shown on Table A2. As suggested by the above formula, multiplying all these estimates and subtracting from 1 would give an estimate of the false negative rate. Where a STATS19 postcode is available, we get 1 – (0.997*0.972*0.976*1) = 0.054 or 5.4%. Where a STATS19 postcode is not available, the equivalent figure is 47.8%.

Table A2 Estimation of false negatives in STATS19-HES linkage (based on 2004 data)

Variable	Maximum tolerance allowed	Est. prob. of agreement for true match	Note
Full or partial STATS19 postcode available			
SHA	Exact or adjacent	0.997	Estimated by matching on exact postcode without using SHA
Age	At most 3 years difference	0.972	Estimated proportion of matches on postcode and sex and age to within 10 years having no more than 3 year difference in age
Date	Date of admission up to 2 days after date of accident	0.976	Estimated proportion of matches on postcode, sex, exact age and date to within 7 days having no more than 2 day difference in date
Postcode		1.000	By assumption; this means the false negative rate is an underestimate
Road user class	*Disregarded (i.e. allow non-match)*		
LAD	*Disregarded (i.e. allow non-match)*		
Estimated false negative rate		***5.4***	
No STATS19 postcode available			
SHA	Exact or adjacent	0.997	As above
Age	At most 3 years difference	0.972	As above
Date	Date of admission up to 1 day after date of accident	0.966	Estimated proportion of matches on postcode, sex, exact age and date to within 7 days having no more than 1 day difference in date
Road user class	Exact match	0.884	Estimated from cases matching exactly on postcode, sex, age and date
LAD	Exact match	0.631	Estimated from cases matching exactly on postcode, sex, age and date
Estimated false negative rate		***47.8***	

Missing or invalid data

The above calculation of false negatives is based on an empirical approach, which allows for some errors in data (for example, estimation of age by the recording police officer), but not for cases where data is completely missing for the four key matching variables (age, sex, region and date).

Table A3 shows the proportion of records on each of the data files with missing data for the calendar period 1999-2007. Assuming that these are independent between STATS19 and HES and thus effectively additive, this suggests that around 4% of matches may have been missed in addition to the false negatives estimated above. No adjustment has been made for this in the calculations presented in Section 5[42], but it should be borne in mind in interpretation.

[42] Note that a calculation of missed matches from the figures presented is not straightforward. Any missing values on the four key matching variables will clearly prevent a match. However, an invalid postcode on the HES file will only prevent a match where there is a valid postcode for the corresponding record in the STATS19 file. Similarly missing values for user class and LAD on HES will only affect the matching rate when there is no valid postcode available. It would of course be possible to exclude these records with missing values from the files used in matching. Whilst this would likely improve the matching rate, it would (for reasons outlined) result in some missed matches.

With the exception of LAD, for which over 10% of HES records in 1999 have missing or invalid values, there are no strong variations in the proportion of missing data over time.

Table A3: Proportion of records on HES and STATS19 files used for linkage with missing data for linking variables (years 1999-2007)

	HES	STATS19		
		Serious	**Slight**	**All**
SHA	0.1%	Compete recording assumed		
Age	0.1%	2.2%	2.9%	2.8%
Date	Complete recording assumed			
Gender	0.1%	0.0%	0.1%	0.1%
Postcode	1.7%	Allowed for in matching		
Road user class	1.2%	Compete recording assumed		
LAD	3.7%	Compete recording assumed		

Summary

This annex sets out one approach to estimating false positives and false negatives for the STATS19-HES linkage, using the 2004 data. If these estimates are representative of the levels over time, the estimated false positive rate over the period 1999 to 2009 would be around 2%, with a false negative rate of around 15% (although allowing for missing data, this is probably higher). These latter estimates would take account of the fact that the proportion of STATS19 records with a valid postcode varies over time.

Annex B: Additional tables and charts (section 4)

Table B1 Linkage results by nature of collision recorded in HES: Casualties in traffic accidents within the scope of STATS19, 1999-2009

Collision type	**Linked**	**Not linked**	**Total**	***Proportion linked***
Car/LGV	101,515	77,882	179,397	*57*
HGV/Bus	8,467	7,552	16,019	*53*
TWMV/3WMV	2,900	3,166	6,066	*48*
Object	16,219	18,142	34,361	*47*
Other motor vehicle	7,855	9,805	17,660	*44*
Pedestrian/Animal	615	1,023	1,638	*38*
Non motor vehicle	321	559	880	*36*
Unknown	14,083	33,362	47,445	*30*
None	16,661	72,383	89,044	*19*
Cyclist	496	2,324	2,820	*18*
Total	**169,132**	**226,198**	**395,330**	***43***

Table B2 Linkage results by road user type recorded in HES for casualties in non-collision traffic accidents, 1999-2009

Road user type	**Linked**	**Not linked**	**Total**	***Proportion linked***
Car (inc three WMV)	8,301	9,146	17,447	*48*
LGV	338	594	932	*36*
HGV	370	682	1,052	*35*
Motorcycle	6,061	17,797	23,858	*25*
Bus	276	1,037	1,313	*21*
Unknown	56	533	589	*10*
Cyclist	1,259	42,594	43,853	*3*
Total	**16,661**	**72,383**	**89,044**	***19***

Table B3 Proportion of HES traffic accident records linked to STATS19 by gender and road user type recorded in HES, 1999-2009

	Female			**Male**		
	Linked	Total	*Proportion linked*	Linked	Total	*Proportion linked*
Car occupant	29,006	59,799	*49*	42,735	82,983	*51*
Pedal cyclist	2,396	15,296	*16*	10,904	56,691	*19*
Motorcycle user	2,526	6,420	*39*	29,439	67,014	*44*
Pedestrian	16,188	28,033	*58*	25,357	46,888	*54*
Other or unknown	2,990	10,762	*28*	7,591	21,223	*36*
Total	**53,106**	**120,310**	***44***	**116,026**	**274,799**	***42***

Table B4 Proportion of HES traffic accident records linked to STATS19 by MAIS level, 1999-2009

MAIS	Linked	Not linked	Total	*Proportion linked*
1	42,734	58,055	100,789	*42*
2	68,654	93,674	162,328	*42*
3	26,472	26,073	52,545	*50*
4	2,523	2,830	5,353	*47*
5	637	653	1,290	*49*
6	1,278	1,283	2,561	*50*
9 (unknown severity)	17,474	25,362	42,836	*41*
99 (not coded)	9,360	18,268	27,628	*34*
Grand Total	**169,132**	**226,198**	**395,330**	***43***

Table B5 Proportion of HES traffic accident records linked to STATS19 by primary diagnosis type, 1999-2009

Diagnosis type	All admissions			Admissions for > 5 days			% admitted >5 days
	Matched	Total	*Match rate*	Matched	Total	*Match rate*	
Organ/internal injury	11,643	25,331	*46*	3,484	7,205	*48*	*38*
Superficial	15,241	34,171	*45*	465	1,061	*44*	*3*
Fracture	84,063	191,475	*44*	28,648	55,099	*52*	*33*
Open wound	17,809	40,729	*44*	1,273	2,613	*49*	*7*
Dislocation/sprain/strain	4,859	13,276	*37*	704	1,483	*47*	*12*
Other or unspecified	35,517	90,348	*39*	2,131	5,201	*41*	*6*
Total	**169,132**	**395,330**	***43***	**36,705**	**72,662**	***51***	***21***

Table B6 Proportion of STATS records linked to HES by casualty type recorded by police, 1999-2009

	Serious			Slight		
Casualty type	**Linked**	**Total**	***Proportion linked***	**Linked**	**Total**	***Proportion linked***
Pedestrian	29,664	67,446	*44*	19,025	261,340	*7*
Motorcycle	25,307	61,866	*41*	11,303	191,743	*6*
Pedal cycle	8,509	24,642	*35*	6,696	155,194	*4*
Car	43,831	134,963	*32*	39,905	1,607,081	*2*
Other vehicle	585	1,890	*31*	382	12,561	*3*
LGV	1,620	5,634	*29*	1,222	55,111	*2*
HGV	992	3,479	*29*	679	22,992	*3*
Bus or coach	718	4,407	*16*	1,013	75,933	*1*
Total	**111,226**	**304,327**	***37***	**80,225**	**2,381,955**	***3***

Figure B1 Proportion of STATS19 seriously injured casualties linked to HES by casualty type and casualty class, 1999 - 2009

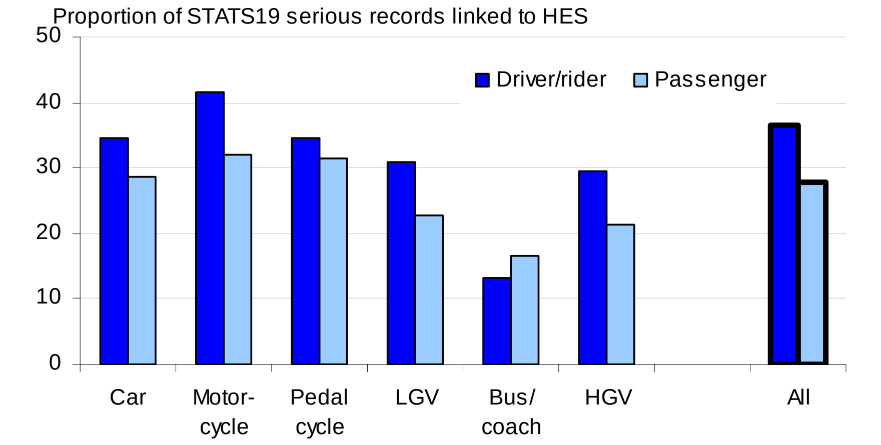


Figure B2 Proportion of STATS19 seriously injured casualties linked to HES by gender and casualty type, 1999 - 2009

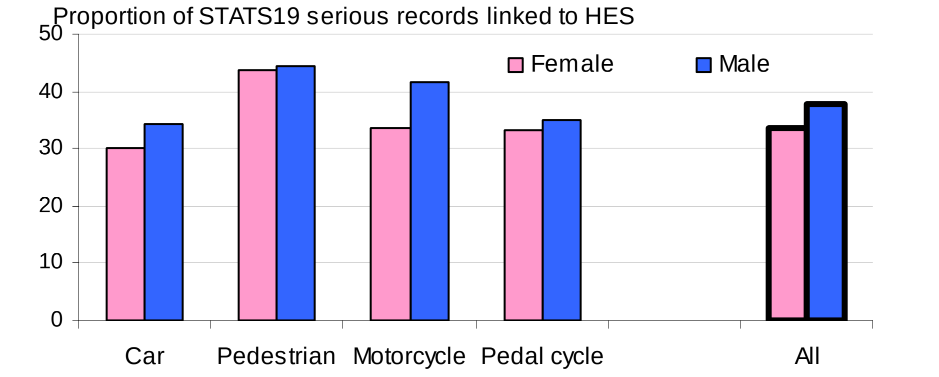


Figure B3 Proportion of STATS19 records linked to HES plotted against proportion of STATS19 with a valid postcode: Police forces in England, 1999 - 2009

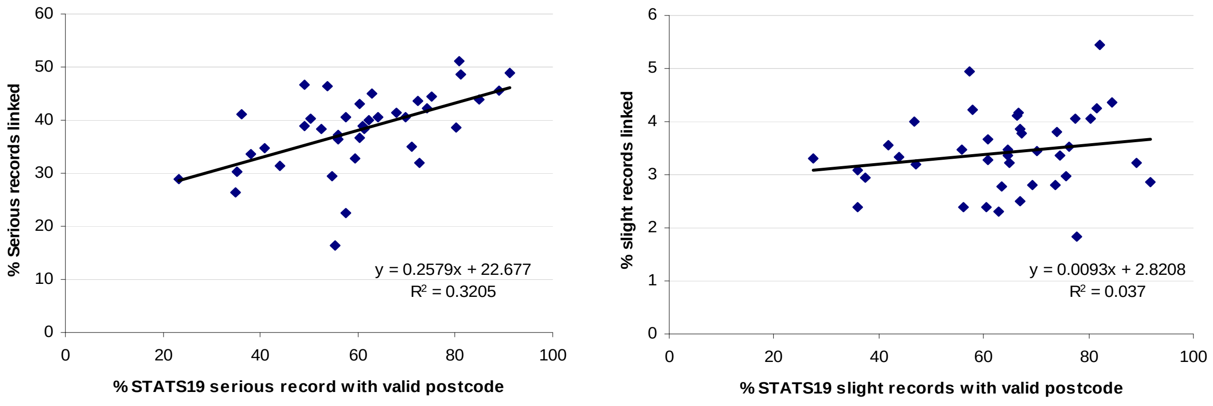


Figure B4 Serious casualties as a proportion of all injuries: STATS19 and linked datasets (indexed, 1999 = 100)

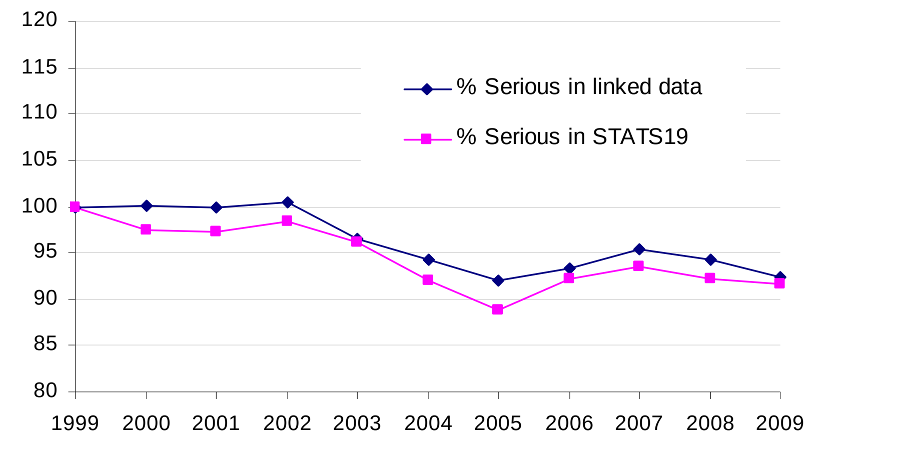


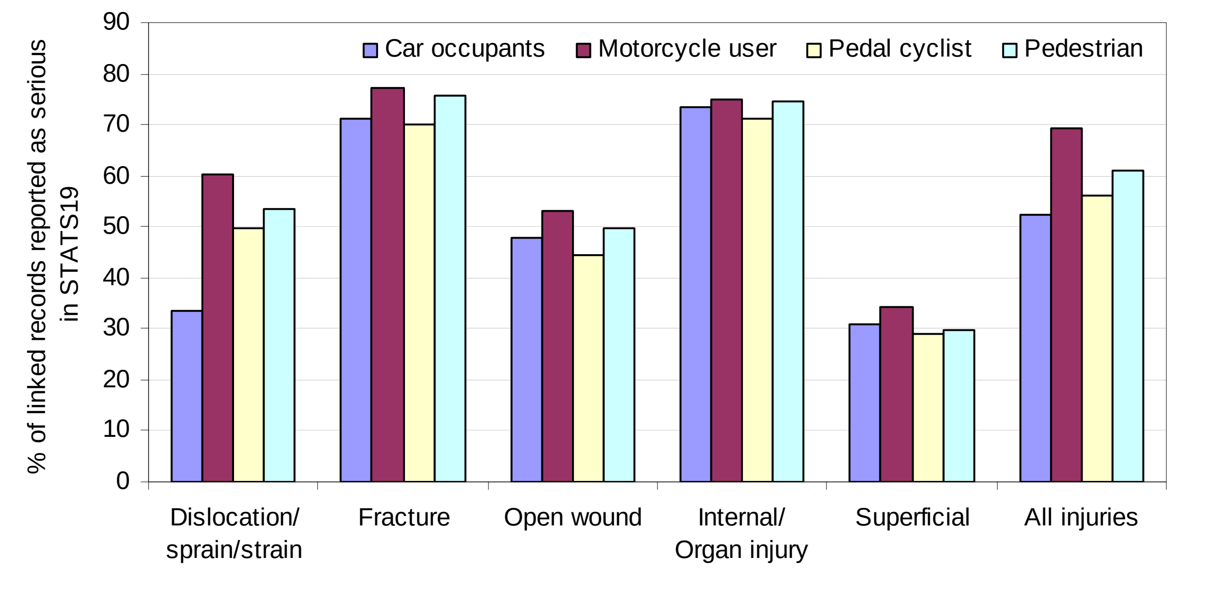
Table B7 Number and percentage of linked records by MAIS/length of stay and STATS19 severity, 1999 - 2009

	Number of linked records			*Percentage of linked records*		
	Serious	**Slight**	**Total**	***Serious***	***Slight***	***Total***
MAIS						
1	19,007	29,897	48,904	*17*	*37*	*26*
2	52,803	24,325	77,128	*47*	*30*	*40*
3	25,053	4,242	29,295	*23*	*5*	*15*
4-6	4,293	578	4,871	*4*	*1*	*3*
Unknown	10,070	21,183	31,253	*9*	*26*	*16*
Total	**111,226**	**80,225**	**191,451**	***100***	***100***	***100***
Length of stay						
0	9,542	20,771	30,313	*9*	*26*	*16*
1	19,755	27,200	46,955	*18*	*34*	*25*
2-4	27,480	16,283	43,763	*25*	*20*	*23*
5+	38,038	9,963	48,001	*34*	*12*	*25*
Unknown	16,411	6,008	22,419	*15*	*7*	*12*
Total	**111,226**	**80,225**	**191,451**	***100***	***100***	***100***

Table B8 Proportion of linked casualties coded as seriously injured in STATS19 data by nature and body region of injury recorded as primary diagnosis in HES, patient admission 1999-2009

	Fracture	Internal/ Organ injury	Open wound	Dislocation/ sprain/strain	Superficial	All selected
Arm/hand	*72*		*42*	*55*	*30*	***65***
Leg/foot	*79*		*53*	*67*	*29*	***74***
Head	*75*	*74*	*48*	*44*	*31*	***50***
Lower back	*71*	*72*	*56*	*60*	*30*	***58***
Multiple	*85*		*45*	*38*	*33*	***59***
Neck	*68*		*53*	*23*	*23*	***37***
Thorax	*61*	*75*	*59*	*33*	*33*	***57***
All selected	***74***	***74***	***49***	***42***	***31***	***58***

Figure B5 Proportion of linked casualties coded as seriously injured in STATS19 data by nature of primary injury and road user type, 1999 - 2009



Annex C: Analysis of linked data: coding of road user and accident type

The linked dataset can be used to assess the accuracy of coding of variables which are available in both the police and hospital datasets, in particular road user and accident type[43]. It is not possible to identify the reasons for any discrepancies (i.e. whether due to inaccuracies in STATS19, HES or the linkage made), but given the nature of the reasons for data collection it seems more likely that these will largely represent miscoding by hospitals, where there is likely to be less interest in the accident circumstances.

Road user type

Table C1 shows overall a good degree of agreement in recording of road user type for linked records. There is clear agreement in 94% of cases where coded in both datasets. This suggests recording of road user type in the hospital data is on the whole reasonably accurate[44] - the table shows that for the main road user groups, nearly 90% the HES coding agrees with STATS19 (which is more likely to be accurate).

Table C1 Distribution of road user group in STATS19 for each road user in HES, linked records with exact agreement on postcode 1999-2009

	STATS19 road user type							
HES casualty type	**Car**	**Goods vehicle or bus**	**Motor-cycle**	**Pedal cycle**	**Other**	**Pedestrian**	**All**	**Number linked**
Car	*89*	*26*	*1*	*1*	*26*	*2*	*44*	**54,051**
Goods vehicle or bus	*2*	*65*	*0*	*0*	*11*	*0*	*3*	**3,900**
Motorcycle	*1*	*0*	*93*	*3*	*14*	*0*	*20*	**24,514**
Other motor vehicle	*0*	*1*	*0*	*0*	*15*	*0*	*0*	**282**
Cyclist	*0*	*0*	*2*	*88*	*4*	*1*	*7*	**8,972**
Animal rider	*0*	*0*	*0*	*0*	*13*	*0*	*0*	**126**
Pedestrian	*1*	*2*	*1*	*4*	*13*	*92*	*21*	**25,469**
Unknown or missing	*7*	*5*	*3*	*3*	*4*	*4*	*5*	**6,617**
All	*100*	*100*	*100*	*100*	*100*	*100*	*100*	
Number linked	**58,141**	**4,307**	**25,298**	**9,081**	**734**	**26,370**		**123,931**

[43] This assumes that the linkage is reliable, and in particular the use of road user type in determining links does not lead to bias. In order to maximise the quality of the linkage for this analysis, it is based on cases where an exact match on postcode was found. For such links, road user type was not used, or less important, in determining linkage.

[44] It should be noted that road user type forms part of the derived 'road user class' variable which is used in establishing linkages, which may introduce some bias to this analysis. However, where there is an exact agreement on postcode, links without agreement on this variable are allowed which should mean that any such bias is minimal and the overall conclusions are generally sound.

Collision type

Table C2 shows the level of agreement in terms of the number of vehicles involved in the collision recorded in HES (collision type derived from the cause of injury coding) and STATS19 (number of vehicles recorded directly). Overall, the degree of agreement appears reasonable. In cases where HES records collision with a motor vehicle and the casualty is not a pedestrian, it seems reasonable to assume that the nearly 3 thousand cases where STATS19 records only one vehicle was involved represent inaccuracy, probably in HES.

Table C2 Number of vehicles in STATS19 by collision type recorded in HES, linked records with exact agreement on postcode for non-pedestrian casualties 1999-2009

	STATS19 Number of vehicles		
HES Collision type	**1**	**2+**	**Total**
Motor vehicle	2,889	53,665	**56,554**
Non-motor vehicle	136	270	**406**
Object/pedestrian/animal	9,863	3,187	**13,050**
None	10,615	4,492	**15,107**
Unknown	2,578	4,096	**6,674**
Total	**26,081**	**65,710**	**91,791**

Coding of casualty class

Finally, table C3 shows degree of agreement in casualty class (driver, passenger or pedestrian) between STATS19 and HES. Again, overall there is a good agreement with 96% of cases recorded as either driver, passenger or pedestrian in HES agreeing with the coding in STATS19 (where the HES coding is known).

Table 4.17 Distribution of casualty class in STATS19 for each casualty class in HES, linked records with exact agreement on postcode 1999-2009

	STATS19 casualty class				
HES casualty class	**Driver or rider**	**Passenger**	**Pedestrian**	**All**	**Number matches**
Driver	*77*	*6*	*1*	*49*	**61,115**
Passenger	*2*	*73*	*0*	*13*	**16,206**
Pedestrian	*1*	*1*	*92*	*21*	**25,469**
Other	*14*	*14*	*2*	*12*	**14,524**
Unknown	*5*	*6*	*4*	*5*	**6,617**
All	*100*	*100*	*100*	*100*	
Number matches	**77,752**	**19,809**	**26,370**		**123,931**

Note that the ‘other’ category shown in the table represents cases where it is not possible to determine the casualty class reliably in HES, but information on road user class is available. In terms of the linkage process, the derived ‘road user class’ variable assigns such cases to the most likely casualty class, base on the road user type. For example, the majority of car occupant casualties are drivers rather than passengers, so where the HES casualty type is ‘car’ and casualty class is not determined, links are allowed with STATS19 driver casualties, but not those recorded as passengers (for bus occupants, the reverse is true as the majority of casualties are passengers). Clearly, this results in some genuine matches being missed. This does not affect those cases where a valid postcode is available in STATS19, as for these records we do not require exact agreement on the ‘road user class’ variable. However, where there is no postcode for linking, agreement on road user class is required. Based on the linkages achieved, this effect alone is estimated to result in at least 5 thousand missed matches. These should be allowed for in the adjustment described in section 5, but this analysis suggests that there may be value in using separate variables for road user type and casualty class in the linkage process rather than combining them, or investigating probabilistic matching methods where missing data can be more easily allowed for.

Summary

The analysis presented here shows that, overall, there appears to be a good degree of agreement in the STATS19 and HES datasets in terms of the recording of road user type and road user class variables, though the agreement is not perfect. It is difficult to explore reasons for lack of agreement in great detail.