RAC Foundation

Mobility · Safety · Economy · Environment

Data Linkage in Road Safety

Bridging the divide to support better health outcomes

Seema Yalamanchili
Imperial College London

February 2024

The Royal Automobile Club Foundation for Motoring Ltd is a transport policy and research organisation which explores the economic, mobility, safety and environmental issues relating to roads and their users. The Foundation publishes independent and authoritative research with which it promotes informed debate and advocates policy in the interest of the responsible motorist.

RAC Foundation
89–91 Pall Mall
London
SW1Y 5HS

Tel no: 020 7747 3445
www.racfoundation.org

Registered Charity No. 1002705

RAC Foundation

Mobility · Safety · Economy · Environment

Data Linkage in Road Safety

Bridging the divide to support better health outcomes

Seema Yalamanchili
Imperial College London

February 2024

About the Author

Seema Yalamanchili is a General Surgeon and a Clinical Research Fellow at the Imperial College London Institute of Global Health Innovation. Her clinical work is based in London, where she has trained in various major trauma centres, and specialises in the care of the severely injured. Her academic background is in global health and major trauma, with her doctoral studies exploring innovative approaches to road collision data. Her work studies both existing and novel data sources and analysis to see how they can augment road safety understanding and intervention, particularly from a healthcare perspective.

About this Report

This report summarises the findings of the Road Traffic Injury – Analytics for Integrated Data (RTI–AID) project which has sought to assess how a range of data sources could be better harnessed to contribute to UK road safety. This included attempting new ways of integrating routinely collected health and transport data as well as ascertaining the potential contributions of a range of new data sources to UK road safety. The project has been delivered by Imperial College London / Imperial College Healthcare Trust and was funded by the RAC Foundation and the Fédération Internationale de l'Automobile (FIA) via their Road Safety Grants Programme. For more information about the project please visit https://www.imperial.ac.uk/centre-for-health-policy/our-work/data-science-and-analytics/road-traffic-injury--analytics-for-integrated-data-rti-aid/.

FIA / ROAD SAFETY GRANT PROGRAMME

During the publication of this report, the Trauma Audit & Research Network (TARN) paused the provision of data to researchers. In what follows, all references to TARN should be taken to include any future source of national trauma registry data for England, Wales and Northern Ireland, whether it be designated as TARN or otherwise. Data for Scotland remains available through the Scottish Trauma Audit Group (STAG).

Acknowledgements

The author of this report would like to acknowledge the support and assistance received from Steve Gooding (Director, RAC Foundation), Dr Elizabeth Box (Research Director, RAC Foundation), Dr Ivo Wengraf (Research & Data Manager, RAC Foundation) and the three RTI-AID supervising investigators: Dr James Kinross, Dr Matthew Harris and Professor Ara Darzi at Imperial College London. Additionally, the author thanks the members of the RTI-AID project steering committee (including Department for Transport (DfT), NHS Digital, Imperial College London Big Data & Analytical Unit, London Ambulance Service, the Trauma Audit & Research Network), and in particular Matthew Tranter (Head of Road Safety Statistics, DfT) and Katherine Williamson (Head of Data, DfT). This research was funded by the FIA and the RAC Foundation. Infrastructure support for this research was provided by the National Institute for Health Research Imperial Biomedical Research Centre (BRC).

Disclaimer

This report has been prepared for the RAC Foundation by Seema Yalamanchili. Any errors or omissions are the author's sole responsibility. The report content reflects the views of the author and not necessarily those of the research funding organisations.

Contents

List of Figures

List of Tables

List of Abbreviations

ADR	Administrative Data Research UK
AIS	Abbreviated Injury Scale
BDAU SE	The Big Data and Analytical Unit Secure Environment
CAG	Confidentiality Advisory Group
CPRD	Clinical Practice Research Datalink
DARS	Data Access Request Service
DfE	Department for Education
DfT	Department for Transport
DHSC	Department of Health and Social Care
DPA	Data Protection Act 2018
DPIA	Data Protection Impact Assessment
GDPR	The General Data Protection Regulation
GPDPR	General Practice Data for Planning and Research
HES	Hospital Episode Statistics
HRA	Health Research Authority
IGHI	Institute of Global Health Innovation
KSI	Killed or Seriously Injured
LAS	London Ambulance Service
MAIS3	Clinically seriously injured (having a maximum AIS score of 3 or above)
NDL	Networked Data Lab
ONS	Office for National Statistics
OSR	Office for Statistics Regulation
PACTS	Parliamentary Advisory Council for Transport Safety
PPIE	Patient and Public Involvement and Engagement
RAIDS	Road Accident In-Depth Studies
RTI–AID	Road Traffic Injury – Analytics for Integrated Data
STRADA	Swedish Traffic Accident Data Acquisition
TARN	Trauma Audit and Research Network
TRE	Trusted Research Environment

Foreword

The publication of police-reported road casualty data (STATS19) stands comparison as a world-leading resource, and has served as a cornerstone of the substantial improvements in road safety we've achieved over the last forty years.

But in a world of big data, computerised cars and medical and scientific advancement it is striking how disconnected and compartmentalised our data on road casualties remains. STATS19-based analysis relies on an improving but still rudimentary understanding of injury and long-term health impacts; medical researchers know little of the physical, mechanical and spatial characteristics of road collisions.

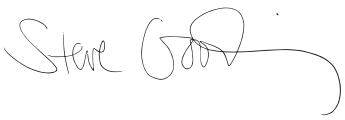
The result is that the road safety community is today confronted with reams of potentially useful data, tantalised with powerful data-science tools for its analysis visible just ahead, but confounded by a suite of convoluted processes firmly planted between where we are now and where we need to be in gleaning useful insights from that data.

Seema Yalamanchili, a General Surgeon and a Clinical Research Fellow at the Imperial Institute of Global Health and Innovation, has been in a unique position in the UK to understand the context, challenges and opportunities in linking data between police and medical sources relating to road safety. The linking of information from multiple sources relating to a particular event or person is the key to unlocking research across a range of disciplines and can in turn lead to a greater understanding of what is taking place at a societal level - invaluable for informing decisions about interventions and policy design.

Are the right road casualties being triaged to the right centres? Why do some demographic groups fare worse than others following similar collision circumstances? What medical interventions can be introduced to improve clinical outcomes for particular injury types? The question for the road safety community is surely not whether road safety-related datasets across the transport, policing and health sectors should be linked, but how?

Ms Yalamanchili set out to establish a deliberately ambitious linkage of four datasets relating to road casualties. Although the requisite approvals were achieved for those linkages the process proved to be extremely challenging. Yes, there are substantial technical, ethical and legal hurdles to be addressed, but Ms Yalamanchili's work shows that they are not insurmountable - lessons have been learnt and solutions documented here for others to build on.

This work offers us an opportunity – as the road safety community strives to find ways to get us back to consistent year-on-year improvements in our road safety performance it's an opportunity that really shouldn't be missed.



Steve Gooding
Director, RAC Foundation

1. Introduction

1.1 Road safety: a global health challenge

Road safety is a global concern. The recent World Health Organisation *Global Status Report on Road Safety 2023* reveals that since 2010, annual road deaths have declined slightly to 1.19 million worldwide.[1] This still equates to more than 2 deaths per minute. In addition, an estimated further 20 to 50 million people will suffer non-fatal injuries, making road safety a leading global health issue.[2]

Europe has some of the safest roads in the world, with the region reporting the lowest fatality rate, at 7 road deaths per 100,000 population.[1] Since the 1970s a number of European countries have consistently been at the vanguard of road safety efforts, notably Sweden, the Netherlands and the United Kingdom. These countries, which have previously achieved significant success, have typically taken a comprehensive approach to road safety with a particular focus on education, legislation, enforcement, post-crash care and an in-depth analysis of collision statistics.[3–5] This 'Safe System' approach builds on established safety principles and evidence from the past 50 years resting on five pillars: safe vehicles, safe road use, safe roads, safe speeds, and post-crash response.[6] Currently, Europe has the greatest concentration of countries with policies and legislation aligning with the Safe System approach.[1]

Whilst significant progress was made within the European Union (EU) between 2000 and 2010, with road deaths declining by 43%, progress since then has stagnated.[7] Neither EU nor United Nations objectives aiming to half road deaths over the decade of 2010–20 were achieved.[8] By 2020, the safest roads were in Norway (17 deaths per million), followed by Sweden (20 deaths per million), then the UK (23 deaths per million).[8] Arguably these nations could be viewed as the victims of their own success, as their already diminished traffic fatality rates make it harder still for them to attain a further significant percentage reduction in road deaths. The European average for reduction in road deaths between 2010 and 2020 was 36%, but in the UK this was markedly smaller at 20%.[1,8]

Nevertheless, with road fatality rates plateauing, the Parliamentary Advisory Council for Transport Safety (PACTS) has stated its disappointment in UK performance, expressing concern that following years of progress the UK had begun to rest on its laurels.[8] This disappointment is heightened by the disparity between current road safety performance and a new goal: Vision Zero. Vision Zero aims to eliminate all traffic fatalities and severe injuries, while increasing safe, healthy, equitable mobility for all.[9] This strategy is increasingly being adopted in Europe, Australasia and North America at both city and national level.[10] Proponents of the Safe System approach and Vision Zero recognise the need for a systematic cross-sector effort, as without this it is unlikely that a comprehensive strategy to tackle the stubborn remainder of road deaths and serious injuries will be found. In addition, addressing concurrent issues such as air pollution, the electrification of transportation and the promotion of active modes of transport such as walking and cycling will also be more challenging to achieve without a comprehensive strategy.

Road fatality rates have been the traditional indicator of road safety progress, but European estimates suggest that for each fatality there are five more people who are seriously injured with life-changing consequences.[8] The *EU Road Safety Policy Framework 2021–2030* sets a new target: to halve total numbers of those killed or seriously injured (KSI). The common definition for seriously injured is casualties with a maximum Abbreviated Injury Scale (AIS) score of 3 or greater (MAIS3+). The AIS is an anatomical injury scoring system based on clinical assessment of the injuries by body region, with each coded according to the AIS dictionary. It therefore requires the use of hospital data rather than police data. Owing to limitations in both hospital and police data, the recommended method is to link the two bodies of information.[11]

Road traffic collision data routinely collected in Great Britain, commonly referred to as STATS19, includes injuries as reported to the police, but these do not constitute formal clinical injury scoring.[12] Over the years, analyses of STATS19 reporting of injuries have suggested both underestimated and overestimated injury burden when compared to the national hospital dataset, Hospital Episode Statistics (HES).[13,14] Both STATS19 and HES have been considered valuable and comprehensive world-leading datasets, but both have also given rise to concerns with respect to their completeness. Furthermore, despite a number of historical linkage projects, there is no routine linkage of HES and STATS19. Where linkage has been performed, there are statistical difficulties with matching records.[15] Today, other nations, many of whom are committing to Vision Zero, such as Sweden,

New Zealand and the United States, are also making concerted efforts in linking road safety data.[16–18] Meanwhile the UK continues to fall behind, steadily losing its position as a leader in road safety progress and innovation.

STATS19 data for 2022 provisionally reported 1,695 fatalities, 29,795 killed or seriously injured (KSIs) and 136,002 casualties in total.[19] Fatalities and seriously injured cases had risen from the lower figures seen during the 2020–1 pandemic period but reduced by 3% from the pre-pandemic figures of 2019. Looking back further however, despite a decline in road traffic injuries overall, there remains a relative plateau in road deaths, which have hovered at around 1,600 killed annually for the past decade. This followed a period of steady decline the decade before.[8]

It is probably too early to say whether this is positive progress or an ongoing rebound from pandemic activity. It is harder still to explain exactly why changes are happening, both more recently and over the previous decade. Although many theories have been put forward – including policing levels, changes in road user type and even governmental focus on Brexit – when it comes to death versus survival on the roads, the role of emergency health services is fundamental and how post-crash care has performed over this time period has not yet been systematically assessed. This component of road safety has to be considered, as over this same time period, significant system changes have been introduced across the nation when it comes to the management of serious injury through the implementation and maturation of regional Major Trauma Networks.

This lack of a data continuum prevents a clear understanding of the rates and nature of road injury by police, transport and health professionals. Without complete data outlining events from collision to injury care and final outcome, those aiming to reduce persistent road traffic collision mortality and morbidity will continue to struggle, often working in silos where transport professionals may draw one conclusion and health professionals another. Fragmented data is no longer acceptable at a time in history where there has been an eruption in digital information, and huge strides in the development of analytical tools to harness insights from them. Without the establishing of mechanisms to join together road safety datasets in a meaningful way, researchers and policymakers will remain locked out of a rich vault of information that already exists but which has failed to be exploited. Without making these connections, breakthroughs in reducing road deaths seem less likely. So the question facing the road safety community is not whether road safety datasets should be linked, but how this could be achieved.

The ‘Road Traffic Injury – Analytics for Integrated Data’ (RTI–AID) project seeks to ascertain what contributions a range of novel data might make to UK road safety, and its implications to road safety data more widely. Run by a research team at the Institute of Global Health Innovation (IGHI) at Imperial College London, and funded by the FIA (Fédération Internationale de l’Automobile) and the RAC Foundation, its main focus has been to assess how emerging big data from various sources may improve our understanding of road safety. The project has two broad research arms: firstly to link together official data from police, ambulance and hospital systems to create a dataset encompassing the casualty’s journey from point of injury to final outcome; and secondly to evaluate the utility of unconventional

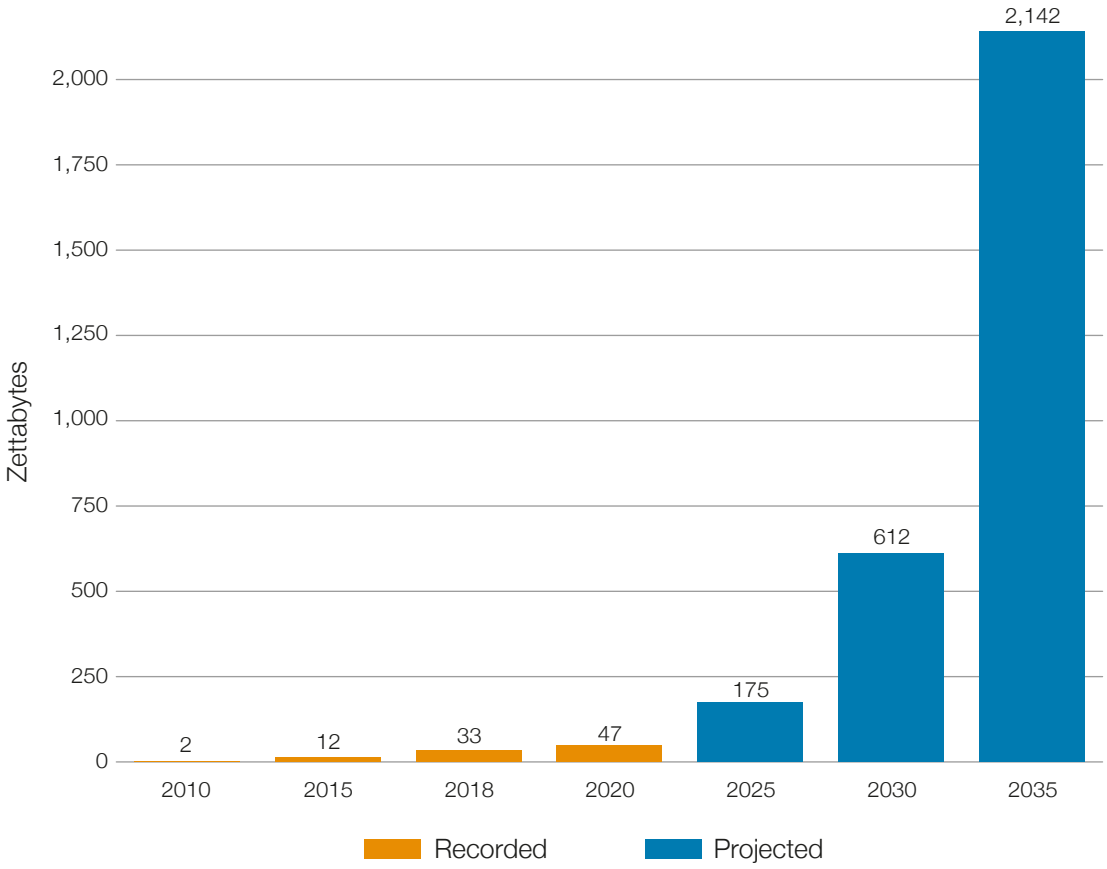
crowdsourced novel data, derived from traffic navigation apps, social media and news media, in identifying road collisions and injuries. This report focuses solely on the first arm of RTI–AID, outlining the findings of work undertaken to link official transport and health datasets. It describes the rationale for undertaking such work and delineates the complex multistep procedures required; but most importantly, it highlights the pitfalls encountered and goes on to recommend a way forward in the hope that this can become a blueprint used by the UK road safety community to robustly establish an optimal linked transport–health dataset, which will prove valuable for future research.

2. The Promise of Linkage in an Era of Big Data

2.1 The era of big data

Recent years have seen an extraordinary explosion in data (see Figure 2.1 and Table 2.1). It is estimated that, during the course of 2023, 120 zettabytes of data will be generated, which is almost 330 million terabytes each day.[21,22] The zettabyte era has affected all areas of society, from industry and retail to government and science. Perhaps some of the most notable applications have been in the health and transport sectors. Big data has remodelled healthcare, predicting disease spread, identifying patients requiring early intervention, and informing health system planning. Within the sphere of transport, a vast amount of geospatial sensor data in particular has contributed to route optimisation, traffic management and efforts in the field of safety. In both sectors new possibilities for research and innovation have arisen.

Figure 2.1: The explosion in global data creation over time



Source: Modified from Statista Digital Economy Compass 2019[20]

Table 2.1: Data volume definitions

Unit	Magnitude	Example
bit	1	A Boolean variable indicating true (1) or false (0)
byte	8 bits	Represents a single character; the base unit of computing
kilobyte	1024 bytes	A long paragraph of text
megabyte	10242 bytes, 1024 kilobytes	A long novel
gigabyte	10243 bytes, 1024 megabytes	17,500 PowerPoint slides
terabyte	10244 bytes, 1024 gigabytes	50,000 trees made into paper and printed
petabyte	10245 bytes, 1024 terabytes	2,000 years of MP3 music
exabyte	10246 bytes, 1024 petabytes	36,000 years of high definition video
zettabyte	10247 bytes, 1024 exabytes	As many bits of information as there are grains of sand on all the world's beaches

Source: Author's own

What counts as big data has evolved over time as the capabilities of analytical tools have grown, but generally the term refers to situations where the scale of data requires significant consideration in processing methods. It has typically been characterised by the three 'V's of volume, velocity and variety (i.e. the rapid generation of vast amounts of information from a range of sources):[23]

- Volume – there are huge volumes of data that require significant processing and storage capability. There is no defined threshold size, with terabytes of data constituting big data for some organisations and petabytes for others, depending on the analytical approach available to them.
- Velocity – the speed at which big data is generated, as well as collected and processed. This is rapid and often takes place in real time. Speeds continue to grow with both web-based content and other mobile and sensor sources, including the Internet of Things.
- Variety – big data began from structured data recorded in relational databases; however, since the birth of the Internet, there has been an even greater semi-structured or unstructured data component including text, images, audio, video, and metadata. This range brings further complexity.

Fundamentally, irrespective of sector, the size and complexity of big data is not of intrinsic value. Value, often considered the fourth 'V', is derived only when data is mined correctly to extract patterns with the acquired knowledge then applied appropriately. It follows that as the volume and complexity of available data grows, the question of how its utility can be unlocked becomes ever more pressing.[23,24]

2.2 Linkage to optimise the value of public sector data

The earliest and most visible application of big data analytics, including data linkage, has been in the corporate sector, particularly the big tech sector, where profit-driven markets have led companies to seek insights about consumers wherever possible, in order to gain a competitive edge. Instances of this are eminently visible in everyday life – for example, the use of Internet cookies to personalise the browsing experience by tracking users and their search history, and thereby targeting advertising to customise their browsing session.

In contrast, despite holding large volumes of digital information, the public sector has been slower to capitalise on the utility of big data. Governments and national agencies generate invaluable large datasets necessary for conducting their core activities, ranging from taxation and policing to national healthcare and transport. Unfortunately, a number of factors, ranging from restricted budgets, concerns over privacy and public perception, the nature of the available skill set and an absence of market forces, have meant that ambitions to extract value from this data have been relatively muted by comparison with the private sector.[25]

Data linkage is the next step to achieving better data utility. Despite the growth in data volume, for policymakers and researchers, single datasets often offer only limited insights and validity. Data or record linkage, also known as entity resolution, is the process of

combining different data sources to create a new enhanced dataset without incurring the additional time and financial costs of further primary data collection. The linking of information from disparate sources relating to a particular entity or individual permits research across a range of different disciplines and greater understanding of what is taking place at a macro level within a population. This can be invaluable for when making decisions about interventions and policy design.[25]

2.3 Mechanistic considerations for linkage

Although not all parties interested in linkage will need to undertake the mathematical process of joining records themselves in depth, a basic understanding of the logic and methods which govern linkage is helpful, as it provides an appreciation of what may or not be feasible in various circumstances.

The aim of linkage is to classify pairs of records according to their match status (i.e. whether or not they belong to the same entity – for example a person, place, institution or event). Developments in computational power and software have increased the feasibility of data linkage, but successful matching of records remains a fraught endeavour, as it is rare that the same unique identifier is present across two datasets. Instead, a combination of non-unique identifiers (i.e. attributes such as – in the case of an entity as a person – name, date of birth or address) and indirect or quasi-identifiers (i.e. not direct attributes but associated events such as a medical procedure or a collision) are employed to create a match (a true pair relationship), a link (an assumed pair) and agreements (noted similarities in the pair).[26,27]

There have been three main methods employed in the absence of unique identifiers: deterministic (rules-based), probabilistic (weighted), and algorithmic. Recently algorithmic methods have also been augmented with machine learning techniques.

2.3.1 Deterministic (rules-based)

The deterministic approach is the most straightforward: it takes a combination of identifiers that should match (i.e. the rule) and, where there is concordance, confirms a match. This is easiest in contexts where a unique identifier is present – for example a national insurance number. Otherwise a specified set of non-unique identifiers such as surname, sex and postcode need to match.

It follows that deterministic methods work well where there are either unique or highly discriminative identifiers, and the datasets are complete and accurate. Since this method depends on unique or discriminative identifiers, there is a low false match rate, but, conversely, it is also prone to missed matches where incomplete or erroneously recorded data impair identifier agreement.[26]

2.3.2 Probabilistic (weighted)

The probabilistic approach can overcome the difficulties associated with incomplete or erroneous data by using a match weight to represent the likelihood of a true match. A range of identifiers are used, each of which are assessed and weighted for any given pair. Each

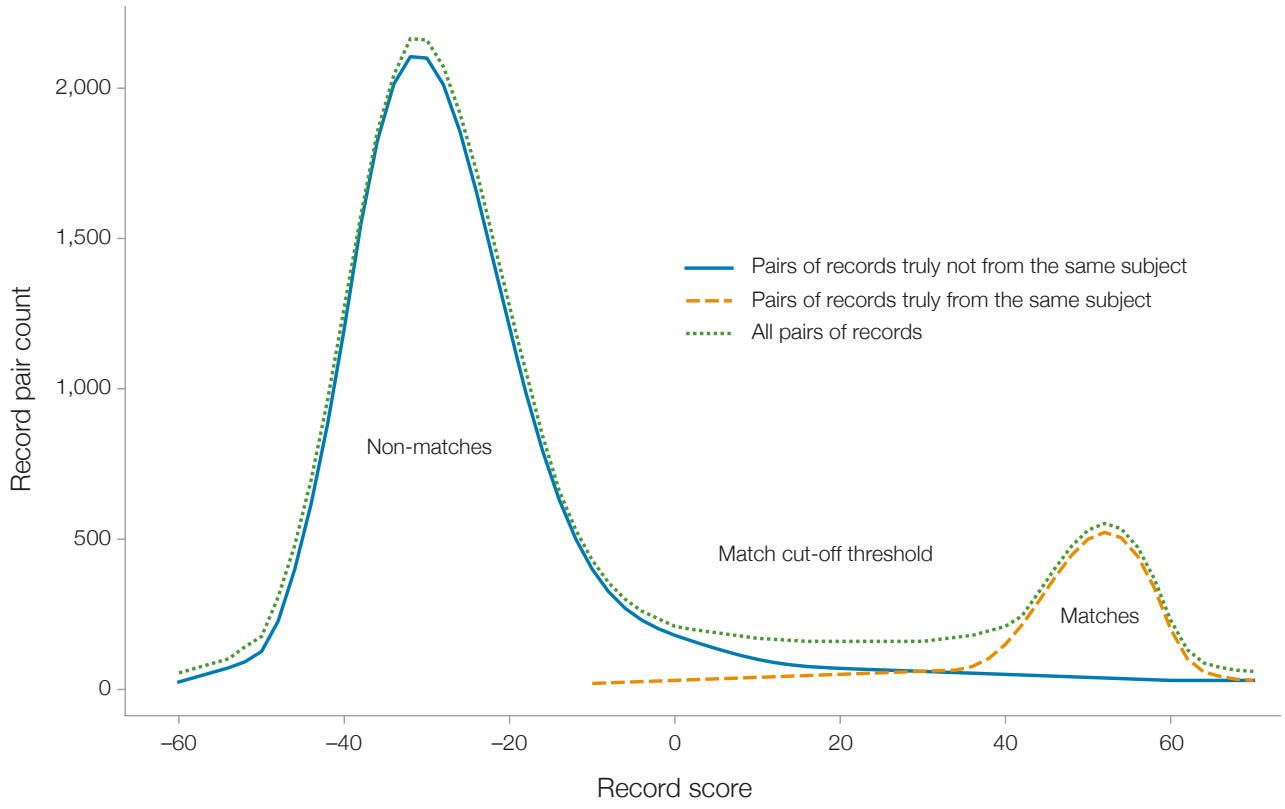
pair thereafter receives an overall weighting dependent on the agreement amongst these various identifiers. A threshold is then chosen, above which any pairing can be considered a link. Figure 2.2 illustrates this methodology.

Various eponymously named statistical models have been used, but the most frequently cited is that of Fellegi & Sunter.[27] In this method, each identifier is assigned a weighting which is based on how discriminative it is – for example, date of birth being more discriminatory than sex. There is a final summation of weights for each pair, and a deduction where there is disagreement in some identifier. Finally a threshold is chosen, and above this cut-off a pair may be classified as a link.

Selecting the ideal threshold is not necessarily easy, and generally requires manual checks and a subjective review. Initially two thresholds are identified: one below which there is clearly no match, and one above which there are undoubtedly true matches. Links in-between are reviewed manually to determine where the final threshold should lie. If training data is available, an alternative to this subjective process is to look at estimated error rates for a range of thresholds and determine the best cut-off from the minimum error rate.

If training or 'gold standard' data is available, it can be used to set the thresholds.

Figure 2.2: Evaluation of match score thresholds in probabilistic linkage



Source: Office for National Statistics, *Developing standard tools for data linkage*[26]

2.3.3 Algorithmic & Machine Learning

In reality a combination of deterministic and probabilistic methods are used, whereby the former is used as a first pass to quickly clear obvious matches and then the latter employed to address cases that are less obvious, and thresholds then considered. Algorithms for linkage can be produced to effect this combinative method, often in an iterative fashion to optimise linkage rate. This process can also be executed through machine-learning methods. Both supervised and unsupervised models, frequently clustering models, are now in use to link data. Although training data improves accuracy, where training data is not available, machine learning may also achieve better linkage than traditional statistical approaches; however, published outcomes for this new methodology are not yet widely available.[26,28–29]

2.4 Balancing rates of matching with linkage error minimisation

Data linkage is a balancing act in a number of respects. Pragmatically, there is a balance to be struck when seeking to maximise the accuracy of linkage, as computational methods and incomplete data both limit linkage rate. Datasets can be incomplete, fragmented and constantly updating. They are frequently not interoperable, and what data is available often requires cleaning to remove duplications, non-standard entries and inaccuracies.

Mechanistically, there is a balance needed to maximise the number of correct matches whilst keeping the number of erroneous links to a minimum. As with other statistical tests applied to data, there is a concession to be made between sensitivity (the proportion of true matches correctly identified) and specificity (the proportion of true non-matches correctly identified), while also taking into account accuracy (the proportion of true matches amongst all matches made). Table 2.2 clarifies these definitions visually.

Deciding thresholds for what constitutes a link or match ought therefore to be decided by data linking technicians working alongside analysts who know the real-world context. The two together are likely to be better able to assess the plausibility of a link, as well as how linked data may influence conclusions drawn from analysis.[26,28–32]

Table 2.2: Definitions of linkage error

Linkage accuracy tool		
Assigned link status	True match status	
	Match (pair from same individual)	Non-match (pair from different individuals)
Link	True match a	False match b
Non-link	Missed match c	True non-match d
Sensitivity (or recall) = a/(a + c); specificity = d/(b + d); positive predictive value (or precision) = a/(a + b); negative predictive value = d/(c + d). In all four parameters, the higher the value the better. In practice, the number of non-matches will usually far outweigh the number of matches, and so the positive predictive value and sensitivity are more informative than the specificity and negative predictive value.		

Source: Harron, British Medical Journal, 2022[29]

2.5 Government and data linkage in the United Kingdom

Across the world, governments and academics have begun to combine data from various aspects of the public sector to facilitate more in-depth population analysis for use by frontline workers and policymakers. There is also a movement to develop data warehouses for linked government data that is reliable and make it more widely available for research.[25,33,34]

Over the past decade, the UK has stated its clear intention for greater data linkage and data warehousing across government and related departments, but progress has been mixed. Following the Digital Economy Act 2017, which provided for greater data sharing between government departments, the Office for Statistics Regulation (OSR) published ***Joining Up Data for Better Statistics*** in 2018, a review which recommended that government analysts prioritise data linkage to improve the production of official statistics.[19] However, in 2021 the Office for National Statistics (ONS) guidance on improving data linkage acknowledged that most linkage conducted across government was often carried out in isolation with little sharing between groups. Further to this, UK efforts toward linkage were not as advanced as those of other countries, particularly those which benefited from using national ID numbers to form links across datasets.[26]

This year, five years after *Joining Up Data for Better Statistics*, OSR published its follow-up report. This report recognised how the pandemic had forced people to use opportunities for digital data sharing, thereby creating a collective change in mindset.[35] However, it also highlighted ongoing doubts as to how data should be shared ethically and legally, and how public concerns regarding data sharing should be addressed. In many ways linkage efforts thus remain at a crossroads.

Although there is still a long way to go to reach the desired scope, availability and utility in public sector linked data, some impressive projects have recently emerged, with collaborations taking place across government, academia and other stakeholders. Table 2.3 furnishes some examples.

Table 2.3: Recent exemplars of public sector data linkage in the United Kingdom

Linkage project	Description	Partners
Education and Child Health Insights from Linked Data (ECHILD) Database[36]	A research database joining existing records in health, education and social care for all children in England in order to better understand the complex relationship between children's health and education, and to identify the challenges that children face and which children are most vulnerable.	Led by University College London, in collaboration with the London School of Hygiene & Tropical Medicine, the Institute of Fiscal Studies, in partnership with the Department of Health and Social Care and the Department for Education, working with NHS England, and the Office for National Statistics.
DATA FIRST[37]	A project to link administrative datasets of civil, family and criminal justice held by HM Prison and Probation Service (HMPPS) and HM Courts and Tribunals Service (HMCTS), as well as data held by DfE, to understand, at scale for the first time, characteristics of returning defendants, outcomes and 'justice journeys'.	Led by the Ministry of Justice in partnership with independent and expert academics.
Connecting administrative vehicle data for research on sustainable transport[38]	Links vehicle attributes and MOT data (DVLA) with vehicle standards (DVSA) for all light vehicles in Great Britain to create a dataset with vehicle type, mileage emissions and location (at the level of LSOAs, Lower Layer Super Output Areas) which can support sustainable transport policy design and local and national level.	Led by the University of Leeds in partnership with the Department for Transport (DfT), DVLA, DVSA, ONS, RAC Foundation, and the University of Bristol.

Source: Author's own

3. Momentum and Trepidation in Health Data Linkage

3.1 Data linkage in healthcare

Healthcare in the UK has been at the vanguard of data linkage through sheer necessity. Efforts to link health data pre-date the wider movement in public sector data linkage, largely as a result of health-related datasets being held by diverse and frequently weakly connected groups, including GP practices in primary healthcare, local doctors in district general hospitals (secondary care) and specialists in regional centres (tertiary care). This is not to mention private groups, non-governmental organisations and charities also delivering various health services. Examples of UK health data linkage are shown in Table 3.1.

Additionally, clinical research frequently hinges on linking record data with other datasets such as patient surveys or basic science work using biological specimens. This means that health data linkage projects tend to be prominent

in the literature and are invariably cited as examples of how to conduct such work.[20,21,30] The USA, Canada, Australia, New Zealand and the UK have all vocally promoted the usage of linked health databases to improve health outcomes, but more recently examples can also be found in the Global South, including South Africa and Bangladesh.[31,40,41]

Table 3.1: Recent exemplars of health sector data linkage in the United Kingdom

Linkage project	Description	Partners
Clinical Practice Research Datalink (CPRD)[42]	CPRD collects de-identified patient data from a network of UK GP practices across the UK. Primary care data is linked to a number of other health data sources to provide a longitudinal dataset representative of the UK, containing records from up to 20% of the population. CPRD has enabled over 3,000 studies covering disease risk factors, drug safety, health policy, and healthcare delivery, over 30 years.	Delivered by the Medicines and Healthcare products Regulatory Agency with support from the National Institute for Health and Care Research, as part of the DHSC.
NHS Digital linked data sets[39]	Formerly the Health and Social Care Information Centre (HSCIC), NHS Digital is the national provider of information, data and IT systems for commissioners, analysts and clinicians for NHS England, providing digital services and managing a large health informatics programme. It has produced Hospital Episode Statistics (HES), a curated dataset covering details of hospital admissions, outpatient appointments and accident and emergency attendance in England since 1989. NHS Digital routinely links HES to a number of datasets including ONS, Mental Health Services Data Set (MHSDS), Maternity Services Data Set (MSDS), Diagnostic Imaging Dataset (DIDS) and Patient Reported Outcome Measures (PROMs).	An executive non-departmental public body of the DHSC.
The Networked Data Lab (NDL)[43]	NDL is a pioneering collaborative network of partners across the UK, established in 2019. NDL takes a federated approach to conduct analyses on linked data from partners, using open analytics, and public and patient involvement to investigate various challenges in health and social care. Projects to date have included COVID-19, youth mental health and unpaid carers. Prior to NDL, the Health Foundation had conducted numerous other data linkage projects including the linkage of ambulance data, children's emergency attendance and primary care for care home residents.	An initiative built and co-ordinated by the Health Foundation.

Source: Author's own

3.2 The ethico-legal demands of health data linkage

The complete process of any data linkage from start to finish, certainly in the UK context, can be conceptually broken down into seven key steps:

1. administrative;
2. data selection;
3. legal considerations;
4. data flow (including storage);
5. privacy considerations;
6. ethical approval and permissions; and
7. mathematical linkage to create the final dataset.

Despite the relative enthusiasm for and progress made in health data linkage, in practice this process involves an additional level of challenge at every stage (see Figure 6.1 in Chapter 6). This additional difficulty when working with health data stems from its inherent nature. The General Data Protection Regulation (GDPR) designates data pertaining to health as a 'special category of personal data', with its own definition and protections. Personal data means any information relating to an identifiable natural person (the data subject), in other words someone who can be identified either directly or indirectly by the presence of an identifier in the data.[44]

'Special category' data is identified by its sensitive or private nature, having the potential to open an individual up to discrimination or interference in their fundamental rights. GDPR aims to reduce the potential for discriminatory profiling on the basis of health data by stipulating specific safeguards for this category of data. These safeguards mean that those undertaking the work have to clearly explain the public benefit of the work; the need for each data field has to be explicit and justified, and data flows must be secure.[45]

In the UK, patients are also entitled to expect an obligation of confidence from their health and care providers. There are some exceptions, such as case reporting for infectious disease outbreaks, but essentially healthcare providers are duty-bound to protect their patients' privacy and not divulge clinical information about an individual without appropriate consent. The National Health Service Act 2006 legally defines the term 'confidential patient information' as information which is given in circumstances where the individual is owed an obligation of confidence, and which contains both identifiers and details of a patient's health or treatment.[46] This definition and those of the GDPR definitions of both personal and sensitive data are shown in detail in Table 3.2.

Table 3.2: Relevant data types and definitions in law[44–46]

Data type	Related law	Definition in law
Personal data	GDPR	Any information that relates to an identified or identifiable living individual. Different pieces of information which collected together can lead to the identification of a particular person also constitute personal data.
Sensitive data	GDPR	Personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs; trade union membership; genetic data and biometric data processed for the purpose of identifying a person; personal data concerning health, sexual life or sexual orientation.
Confidential patient information	Section 251 (11), National Health Service Act 2006	Information about any patient, alive or dead, that meets the following three requirements: (1) is identifiable or likely to be identifiable, for example from other data likely to be held by the person or organisation receiving the data – if a patient could be identified from it (2) was given in circumstances where the individual is owed an obligation of confidence (3) conveys some information about the physical or mental health or condition of an individual, a diagnosis of their condition, or information on their care or treatment.

Source: Author's own

To remain within the confines of the law, some health researchers use aggregated anonymised or pseudonymised datasets (whereby personal identifiers are replaced by artificial identifiers, or pseudonyms), but for linkage projects this is not usually possible. Reasonable match rates depend on the presence of a range of unique identifiers, and non-unique identifiers or quasi-identifiers, and so a balance has to be struck between obtaining and using highly identifiable sensitive personal data and high-quality linkage.

The law aims to protect the rights of the individual by providing two ways in which health researchers are permitted to use their identifiable health data. The first is through informed consent (i.e. the data subject must give specific permission for the research to be undertaken on their data) – however, this is generally not feasible. The datasets are usually historic, even if only a few months old, and constructed through routine collection which does not involve specified consent for the various health research projects which might be conducted on them. Obtaining retrospective consent from the thousands or even millions of data subjects for large-scale database research and linkage is impractical. The alternative is to apply to the national Confidentiality Advisory Group (CAG) to obtain special permission to temporarily suspend the duty of confidentiality for a specific purpose, providing that CAG concur that the benefits of the work can justify the temporary break in confidentiality. Whilst possible, the research group must first engage with the public to establish positive support for the planned linkage work.[46–48]

3.3 Politics, privacy and confidence in health data

Although there has been some progress and success with linkage of health datasets in the UK, there have also been missed steps along the way. It is important to reflect on these, given that much of the ethico-legal underpinnings of most health linkage work centre on how the public view the sharing of their sensitive data. There has to be public support for the use of their private information, and that in turn is dependent on transparency about how it will be used and why.

Perhaps the most notable of these linkages early on was the care data programme set out by NHS England in 2013. The programme set out to link primary (GP) and secondary (hospital) care datasets, in a much-needed effort to connect patient data so as to make it available wherever a patient accessed the NHS. Despite compliance with legal requirements for data sharing, and initial support from the medical community and patient groups, a poor campaign to explain the programme failed to gain public confidence. A couple of data breaches occurring shortly afterwards, and deep concern that data would be shared with commercial companies, obliterated trust amongst doctors and patients alike. Over one million people applied to withdraw consent.[48] After a number of postponements, the scheme was finally scrapped in 2016.[49,50]

The Caldicott Report by Fiona Caldicott, the then National Data Guardian for health and social care, had advised that there was a need for better technical standards, an explanation and promotion of the benefits, an easy opt-out and a dynamic consent process. Although not heeded at the time, these principles remain the foundation of successful patient and public engagement to this day.[51]

Regretfully, almost a decade later, these central tenets are still not well executed. In May 2021, the Government announced the General Practice Data for Planning and Research (GPDPR) programme. This was to facilitate the flow of pseudonymised data from GP surgeries in England into a central NHS database, for purposes other than direct healthcare. Although envisaged as a means to aid more rapid and effective health research and planning, again concerns about transparency, re-identification[1] (and poor public engagement led to millions opting out of the scheme. Furthermore, on this occasion, both the British Medical Association and the Royal College of General Practitioners jointly also expressed their concerns at the general lack of transparency, and of public information and education efforts which undermined the possibility of informed consent. In a development that faintly echoes recent history, GPDPR has been put on hold until more robust processes can be put in place for public engagement, data opt-out or deletion and analysis.[52]

1 Re-identification refers to an instance of anonymised personal data being matched with the subject to which the data refers, thus re-establishing the deliberately severed relationship between the two.

Following the COVID-19 pandemic, public trust has been further eroded with fears over the role of foreign private firms in the acquisition and use of digital patient data, and in the development and maintenance of digital infrastructure for the national health service. Privacy groups have been critical of these companies for their part in designing and running a number of services, including the NHS COVID-19 contact-tracing app with its associated access to confidential data.[53] More recently there has been widespread concern regarding the lack of transparency in the allocation of contracts to build NHS data platforms for storing and processing confidential data, all of which feeds into a sense of mistrust.[54] Hence any efforts to use patient data for linkage must follow Caldicott's advice and make its agenda, processes and safeguards as clear as possible to engender trust and support from the public.

4. The Imperative for Establishing Linkage in Health and Road Safety

4.1 Road safety – the wider story

Road traffic safety encompasses all measures taken to reduce or prevent the injury or death of road users, maintaining that collisions are not inevitable and that in the event that they do occur, injury or death can also be avoided. There has been a shift from considering road collisions as 'accidents', implying that they are unfortunate and unpreventable events, to now noting them as 'collisions' with a multitude of contextual risk factors.[55,56] Although this change in language has taken some 20 years to enter mainstream terminology, structured approaches assessing how and why collisions and injuries occur date back as far as the Haddon matrix in 1970 (see Table 4.1).[57]

Table 4.1: The Haddon Matrix for injury prevention

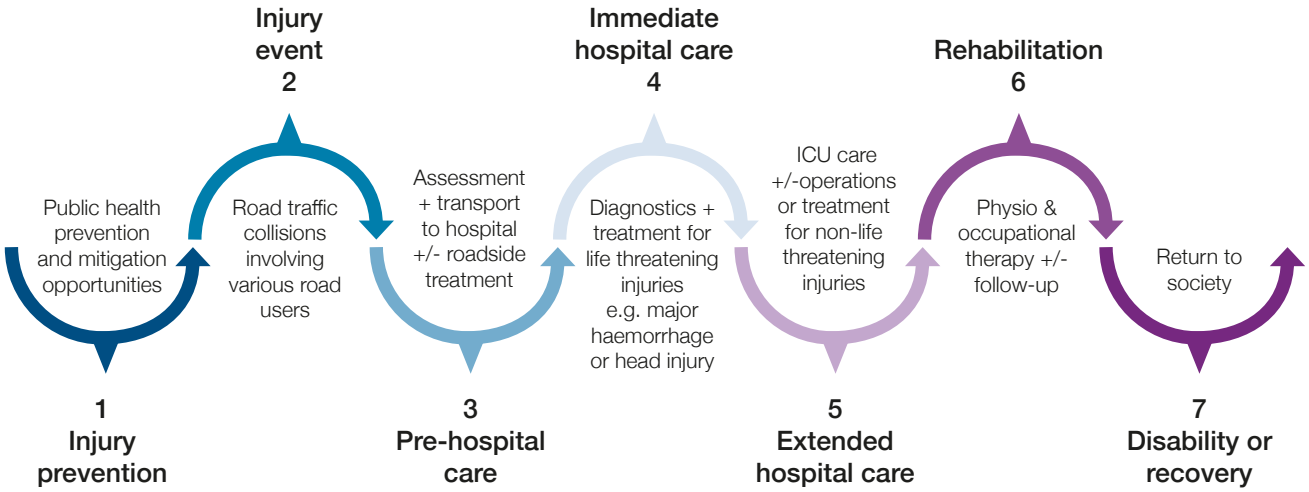
Haddon Matrix				
	Host	Equipment	Physical environment	Social environment
Pre-event/crash				
Event/crash				
Post-event/crash				

Source: Haddon, American Journal of Public Health, 1968[57]

The matrix lists the components of a collision (human factors, vehicles and the environment), and maps them against the phases of a collision (pre-crash, crash and post-crash) to create a comprehensive profile of road injury causation. Although the matrix has been surpassed by other more comprehensive and dynamic systems-based models for studying road collisions,[58] Haddon's contributions remain notable for two reasons. Firstly, Haddon's solution-oriented matrix was embraced well beyond the study of road safety, and was absorbed more broadly in the field of injury prevention as a whole, where it continues to have an impact on public health approaches.[58] Secondly, William Haddon himself was an exemplar of what innovative solutions can be generated when there is cross-fertilisation of ideas across distinct disciplines. Haddon, often considered the father of road safety, trained at both the Massachusetts Institute of Technology and Harvard Medical School, and went on to be both a physician and head of various highways agencies in the United States.

This true multidisciplinary approach to road safety is not as common as it should be, even today. Much investigation, research and campaigning is relatively siloed – often for understandable reasons ranging from cultural and political to practical and financial. Traditionally, there has been a strong presence from transport, engineering and policing disciplines but overall less representation from the healthcare sector. A road crash may implicate a range of organisations from police and transport agencies in its early phase, through to pre-hospital or ambulance groups in the intermediate phase, and finally hospitals and rehabilitation. Even in the health sphere, those in public health studying road collisions rarely enter a discourse with the clinicians treating the injuries on the ground. Whilst it may not be necessary for every one of these professional groups to speak with each of the others at every turn, an increased cross-fertilisation of ideas, to develop a holistic approach to investigation and integrated thinking behind policy design would be beneficial. Figure 4.1 sets out the stages of an injury caused by a collision, starting with pre-crash prevention.

Figure 4.1: The timeline of road injury from prevention, to collision, to healthcare

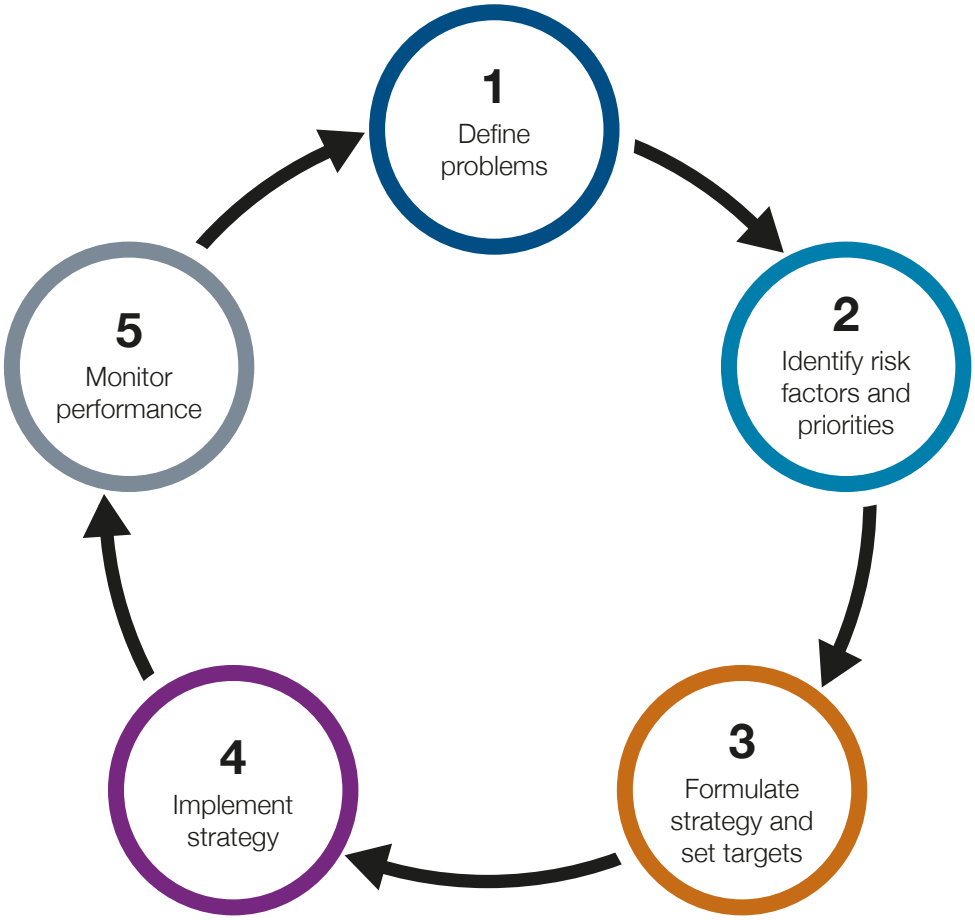


Source: Author's own

4.2 The importance of road safety data

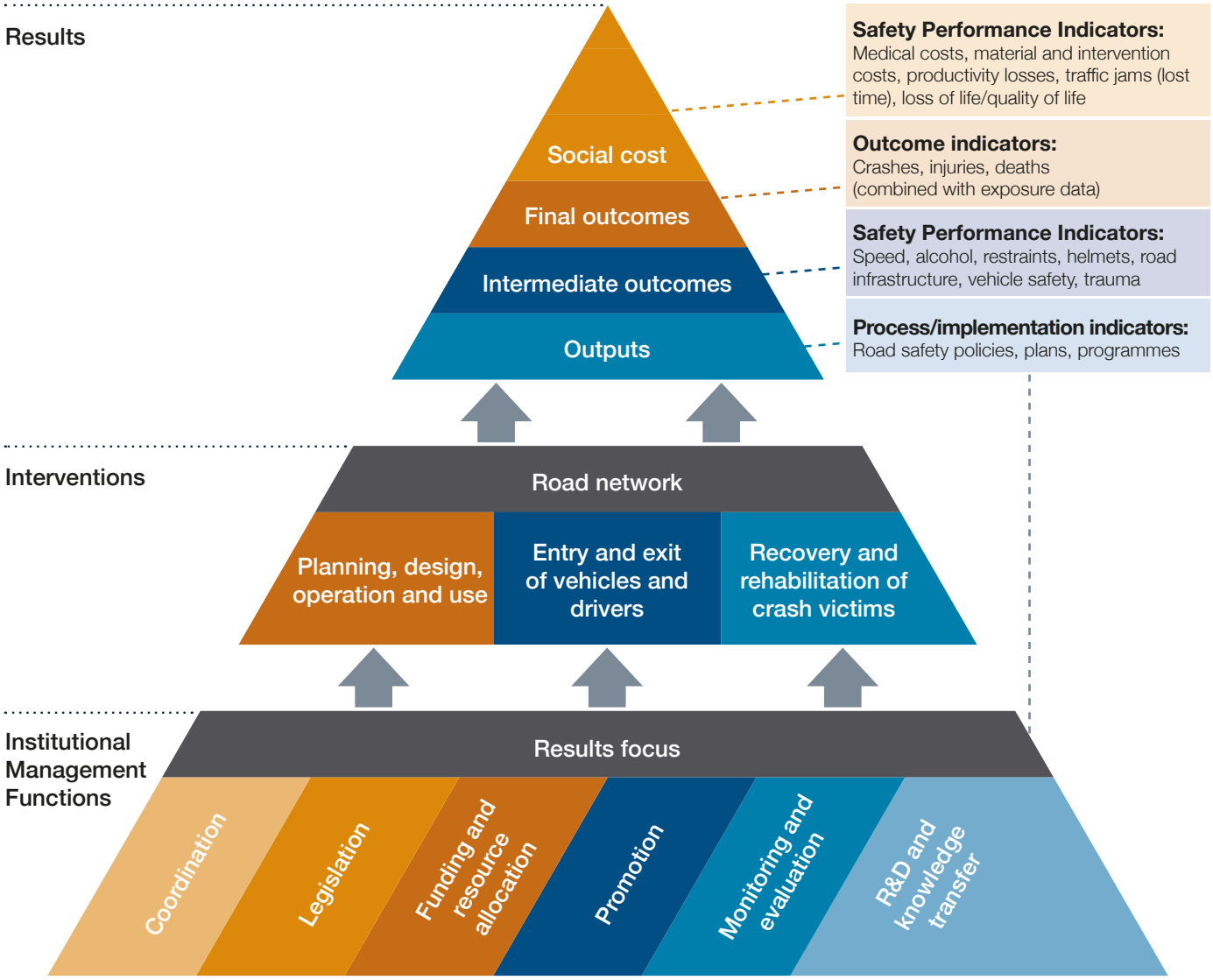
Accurate road safety data is a prerequisite for effective road system management. This data needs to capture details pertaining to causal factors, the nature of the incident itself, and the outcomes. The World Bank has provided guidance on what constitutes a minimum dataset.[59] Furthermore, data should also be complete in terms of covering all incidents to avoid creating a skewed picture. Without this data an evidence base cannot be built. Having evidence is essential in order to characterise the key problems, prioritise risk factors, target interventions and evaluate their progress. In turn, this evidence can also be used to justify funding and raise awareness. The circular nature of good road safety policymaking, using feedback, is illustrated in Figure 4.2. The World Health Organization and World Bank have also conceptualised the road safety management system as a hierarchy of three elements: institutional management functions, which generate interventions, which in turn create desired results.[59] Each of these three elements needs to be assessed, but the linkage of data across these elements is equally important for improved road safety performance (see Figure 4.3).

Figure 4.2 The policymaking cycle for road safety



Source: World Bank 2021 & World Health Organization 2010[59]

Figure 4.3 Overview of road safety data and its function for evidence-based management



Source: World Bank 2021 & PIARC 2019[59]

4.3 Challenges to road safety data

In reality, complete and comprehensive data is difficult to attain. For many countries, road safety data collection systems are still nascent and there is less legal/cultural impetus to report collisions and outcomes. Data may be reported only in the case of severe injury or death and if so, such outcomes may be reported only at hospitals or mortuaries, rather than by police or transport authorities who also log causal factors.[59]

Even in countries where it is mandatory to report motor vehicle collisions, not all events are reported. In most cases collisions are reported to the police, who are well placed to attend road incidents and assess outcomes, but are often under pressure to carry out other responsibilities. In other cases, state transport departments are responsible for

data collection. In Europe, where many nations feel that their police road safety data is accurate, problems remain concerning the issue of completeness of the data collected. Greece, Poland and the Netherlands have all published findings relating to discrepancies in mortality rates between police and hospital data, with police data missing 5–25% of deaths.[60] A number of potential underlying causes of this mismatch have been put forward including the police inevitably being unable to attend all collisions, staff shortages, and public underfunding reducing the ability to complete administrative tasks; to this can be added public disinclination for police attendance or reporting, and differences in definitions for record keeping.

Some collision types, even non-fatal, are also likely to go under-reported, including those involving only a single motorised vehicle, minor injuries, or active road users such as cyclists, and those affecting pedestrians only.[59] As active travel is increasingly encouraged, capturing the reality of safety for these road users becomes increasingly important.

4.4 Linkage of road safety data

As with other complex problems affecting a large population, road safety information is not housed in one single dataset. Although information reported to the police is the cornerstone in most instances, additional datasets are invaluable for a variety of reasons. From a transport perspective, cross-correlation with other information on overall vehicle types and road user numbers is invaluable. Total road user numbers permit calculations of the rate of particular kinds of collision. Similarly, total mileage travelled by different road users allows assessment as to whether collisions are happening because there are many vehicles, or fewer vehicles being driven a lot. From a human cost perspective, corroboration regarding health outcomes with health datasets from hospitals, death certificates and national statistics can confirm the nature of more severe or fatal injuries in particular, and help monitor the quality of official road injury reporting.

Swedish Traffic Accident Data Acquisition (STRADA) is Sweden's road safety data system, comprising crash data collated by the police and health data from patients attending hospital. Information exchange between these two datasets generates a clearer understanding of the actual health outcomes, but covers only those patients with serious injury and only in cases where the patient has consented to their health information being released.[16]

4.5 Efforts to link road safety data in Great Britain

In Great Britain, national road statistics are derived from data collected by the police and collated by the Department for Transport (DfT), in a national database termed STATS19. STATS19 data is collected when a public highway collision involving at least one vehicle (motorised or non-motorised) results in an injured party. It is considered a relatively rich dataset pertaining to the circumstances of the vehicle collision, although these documented contributory factors are initially based on subjective assessment and may not be corrected until the full investigation is complete. Similarly there is relative weakness in recording

outcomes.[61] Although the police do record injury codes as part of data collection, these are still based on a subjective assessment by someone who is, in medical terms, a layman. Thus the documented injury codes in STATS19 do not always align with any formal clinical diagnoses made by clinicians. Most commonly these are fully elucidated only once a full assessment has been made in a hospital.

There have been a number of efforts to link STATS19 to health data to better understand circumstances leading to particular injuries. However, there is no shared unique identifier between STATS19 and health data, and so there has been much variation between each of the different linkage projects, each having different aims, mechanisms and findings.

4.5.1 STATS19 to HES

The longest established linkage is between STATS19 and HES, and was performed initially for data from the period 1999 to 2009. Linkage was found to be possible for 32% of HES records and 37% of STATS19 records, although a proportion of missed matches were assumed. Of HES records within the scope of STATS19, 41% were actually linked to STATS19, rising to 47% when non-collision cycling incidents were excluded. The linkage methodology relied upon casualty age, gender, road user type, casualty residential postcode, local authority district of collision, date of collision and local authority district of the hospital.[62] The linkage was periodically repeated by NHS Digital but over time the rate of successful linkage varied.[63] This has been due in part to changes in healthcare structures such as the creation and development of Major Trauma Networks. These networks protocolled which types of patient were triaged to which hospital on the basis of roadside paramedic assessment and the degree of injury, rather than simply triaging to the closest hospital.[64] Therefore any linkage methodology based on geography and assuming conveyance to the most local hospital would be undermined. Involving healthcare professionals with major trauma system expertise in designing the methods for linkage could improve success in data record matching and interpretation of the findings.

4.5.2 STATS19 to TARN

More recently, DfT has conducted a feasibility study looking at linkage of a subset of HES and Trauma Audit and Research Network (TARN) data in the East of England for 2018–20.[65,66] TARN data was limited to sex, age, time, date and the location of the incident precipitating admission, admission time and date, and finally outcome (death, critical care length of stay, and overall length of stay). Both rules-based and probabilistic weighted approaches were tested with similar results using incident date, time, casualty age and sex. Owing to the sensitive nature of postcode data, this was not initially included. Overall, 43% of TARN records had at least a probable match to STATS19, rising to 62% for TARN records containing incident location data. When the TARN data was filtered so as to limit it to casualties most likely to be in the scope of STATS19 data, the linkage rate rose to 68%. Later, using a small subset of data with the first part of the casualty home postcode provided increased the linkage rate to as high as 87%.

This is an important illustration of how, in the absence of a unique identifier, the use of one or two key variables can substantially boost linkage rates. However, these particularly useful

non-unique but discriminatory identifiers tend to be more sensitive in terms of confidentiality, so there needs to be a considered justification for the use of such data fields to optimise linkage rates. Depending on the sensitivity of the identifier, and other features in the study design, a decision to use it may be reasonable in order to sustain meaningful linkage and avoid undermining the overall aims and potential benefits of the work.

Further smaller-scale linkage projects have also linked STATS19 to TARN, but this was mainly for specific road user groups. These include the Targeting Road Injury Prevention Project, which looked at risk factors in driver culpability in the killed or seriously injured groups (KSI), and DocBike, focusing on motorcyclist collisions. Again, geographically, these were restricted to certain regions only. [67–69]

4.5.3 Road Accident In-Depth Studies (RAIDS)

Since 2012, using a more intensive and concentrated approach, the Transport Research Laboratory has conducted the RAIDS project, investigating a select number of collisions. Each of these collisions is explored in depth, delving into information from both the collision investigation and the associated casualty hospital data. This cross-referencing analysis shows how particular collision factors can lead to specific injuries.[70]

Recently, the RAIDS database has been used by university researchers to investigate the link between the physics of collisions and brain injury. The AutoTriage project established how the vector of impact (higher lateral change in velocity) influenced more severe brain injury, as well as the efficacy of cycle helmets in reducing mild to moderate neurological injury, irrespective of impact vectors.[71]

In all cases, the linkage work has been illuminating, providing a greater understanding of the interrelationship between collisions and injury outcomes, but more could be asked of this data if robust linkage were established as standard.

5. The Road Traffic Injury – Analytics for Integrated Data (RTI–AID) Linkage Project

5.1 RTI–AID rationale and aims

The Road Traffic Injury – Analytics for Integrated Data (RTI–AID) research team, based at the Institute of Global Health Innovation (IGHI) at Imperial College London, was formed from a combination of health disciplines including clinicians working at St Mary's Hospital, a leading London Major Trauma Centre, as well as global and public health specialists with an interest in major trauma and health systems. This academic healthcare perspective was influential in generating the objectives and design of the project, but a wide range of stakeholders from the road safety community were assembled as a Steering Group which was consulted at key stages.

RTI–AID sought to assess how a range of data sources could be better harnessed to contribute to UK road safety. This included attempting new ways of integrating routinely collected health and transport data as well as ascertaining the potential contributions of a range of new data sources to UK road safety, and the implications of this more widely.

With respect to traditional road safety datasets, those arising from the transport and health sectors, it specifically aimed to perform a definitive linkage that would describe road injury from the point of collision until hospital treatment and discharge. This goal was chosen for a number of critical reasons:

- each of the practitioners (police, ambulance paramedics and hospital clinicians) attending to road casualties play an important role and collect equally important and relevant information, but hitherto these datasets have not all been linked to piece together a holistic picture;
- use of the different datasets could not only add to more understanding of what happens to each casualty, but additionally might be able to capture a greater number of casualties in total than the sum of those from each individual dataset;
- given the absence of a shared unique identifier across datasets, the use of the pre-hospital or ambulance dataset held out the possibility of contributing to higher linkage rates, especially for the more seriously injured patients, as this data describes conveyance from collision to treating hospital; and
- the completed linked dataset could not only provide the most comprehensive description of the road injury population to date, upon which new health and road safety research could then be conducted, but also had the potential to be used to verify novel emerging sources of road safety data, such as navigation apps, and news and social media reports.[72,73]

5.1.1 Datasets

RTI–AID therefore set out to establish an ambitious linkage of four datasets, focusing initially on the Greater London Area as a pilot. These four datasets were STATS19 (collected by the police, curated by DfT), ambulance data (collected and stored by London Ambulance NHS Trust) and hospital data in the form of HES and TARN. These latter two were both chosen for inclusion because of the way they complement each other: whilst HES should capture most injured patients, it provides less detail on clinical outcomes, while conversely TARN captures only the more severely injured but at the same time provides a more detailed description of actual injuries and interventions (see Appendix A). The principal features of each of these datasets may be compared in Table 5.1.

Table 5.1: Features of road safety datasets for linkage

Database	Collection	Data holder	Inclusion criteria	Data fields	Lag to release	Notes
STATS19	Collected by and reported to police	Department for Transport	Road traffic collisions with at least one casualty	Time, location Road user types Contributory factors	September of subsequent year	In London, collated from various police forces by Transport for London
London Ambulance Service NHS Trust	Recorded by paramedics in real time	London Ambulance Service	All incidents logged as a road traffic injury on arrival	Casualty demographics Alert and transport times	Three months later	Clinical data also available but not in accessible digital format
Hospital Episode Statistics (HES)	All hospitals	NHS Digital	All attendances to A&E and hospital admissions	Date and time of admission Injury codes	Provisional: one month later; final release: one year later	Not designed to record nuances of injury; incomplete data capture at hospital level
Trauma Audit and Research Network (TARN)	Hospitals designated Major Trauma Centres & Trauma Units	TARN	Patients with specified injuries and also three-day admission / ICU stay / death	Date and time of admission Injury-specific clinical details	Three months	Predefined injuries for inclusion do not necessarily capture all the seriously injured

Source: Author's own

5.1.2 Geography

The geospatial bounds of the study were limited to the Greater London area to limit costs, as this was an experimental study – but this was also important for other practical reasons including the geographical overlap of the London Major Trauma Network. The network includes the NHS pre-hospital (ambulance and Air Ambulance services) and hospital services caring for injured patients in Greater London, and is the best established trauma network in the United Kingdom. It was hypothesised that by restricting the focus to within this region, injury care would be relatively comparable, or at least the findings more easily interpretable.[74]

Furthermore, whilst STATS19, HES and TARN are national, at the inception of the study, ambulance data in the United Kingdom was available only via individual regional ambulance NHS trusts, such as the London Ambulance Service (LAS). Therefore limiting the study to Greater London pragmatically also limited ambulance data to one source with one data controller.

5.1.3 Outcome

Following a series of discussions and iterations, a successful study design with deliberate decisions about data selection and data flow was created and approved by the relevant data controllers. It received patient and public support, and ultimately the linkage received ethical approval by the Health Research Authority (HRA) and CAG.

Unfortunately, pandemic-related pressures exacerbated by personnel changes, along with time and budget constraints, all led to the halting of the work prior to data flow for the linkage, but the work conducted to date nevertheless represents a blueprint for study by those seeking out further linkage attempts. This outline together with a number of learning points garnered from the experience describes how the road safety community could attempt to reproduce a similar linkage in the future.

5.1.4 Learning points

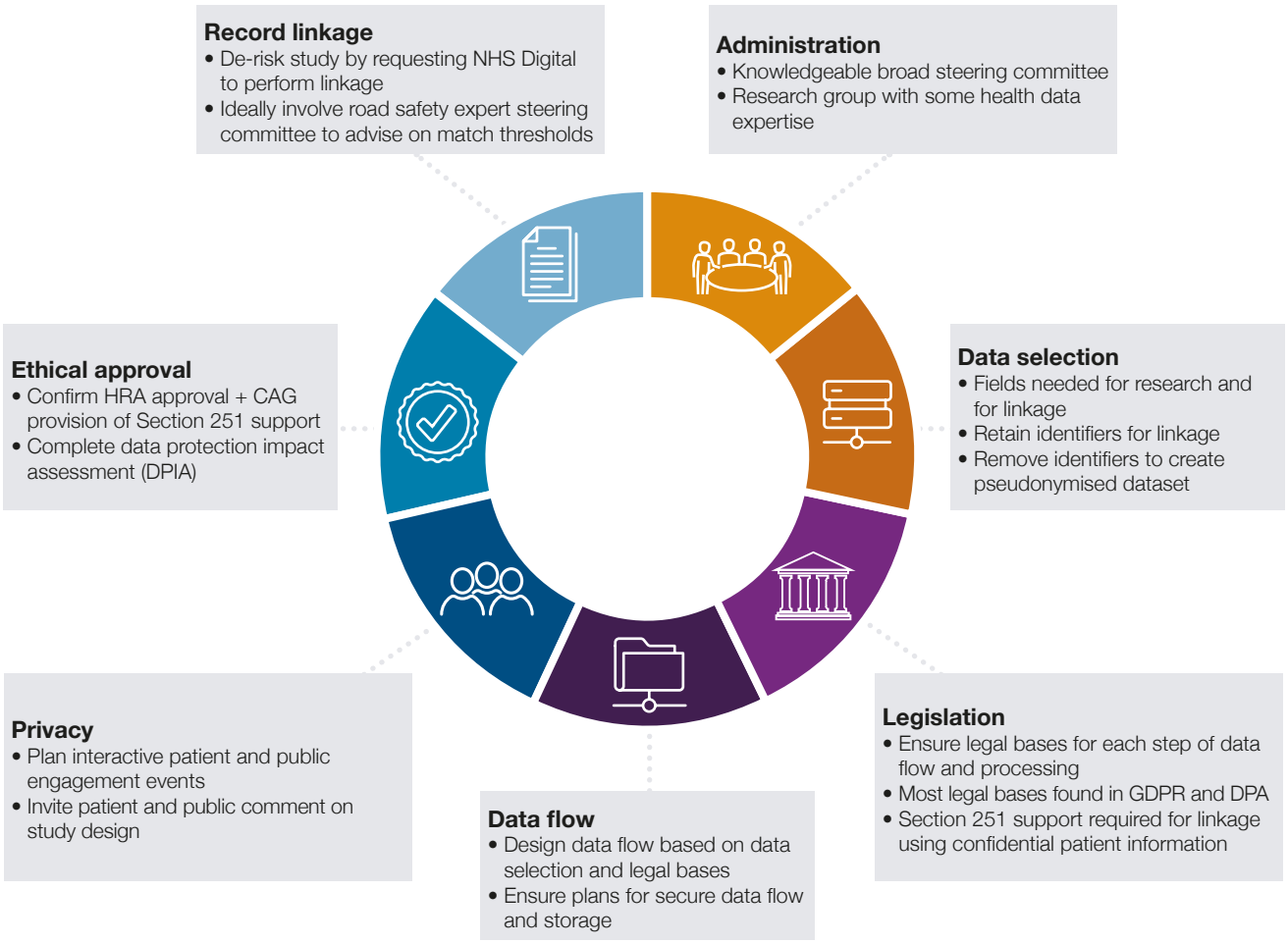
Although the RTI–AID project received the requisite approvals for the data linkage of a range of road safety datasets, this process was extremely challenging and proved to be a constant learning process throughout. The linkage of four major public sector datasets, each owned by different organisations in two separate sectors, was a significant administrative undertaking. With most previous linkage efforts being conducted between only two datasets, in this undertaking the usual process was effectively tripled. There was also an explicit aim of obtaining the best data for both linkage and analysis, so there was no strategy for avoiding complex ethical approvals, which made for an even greater administrative burden.

Many of the steps required were not clear at the outset, and were instead identified and addressed as they arose, with the project plan being altered and refined along the way. Significant learning points for the wider road safety community should be derived from the experience, both from understanding the necessary administrative steps and also the less visible smaller processes and hurdles encountered.

6. RTI–AID Linkage: Study Design, Hurdles and Lessons Learnt

The following section outlines the steps undertaken in the data linkage of established datasets collected by police, ambulance and hospitals within a Major Trauma Network. This section is set out using the seven key steps in establishing health data linkage: administration, data selection, legal considerations, data flow (including storage), privacy, ethical approval and permissions, and record linkage (see Figure 6.1). For each of these steps there is an explanation of the RTI–AID design and related discussions, together with the related barriers and learning points encountered over the course of the project.

Figure 6.1: Key processes for conducting health data linkage



Source: Author's own

6.1 Administration

Linkage of large datasets is rarely confined merely to the statistical means of establishing matches or links, and this is particularly so in the case of health data. There is specific regulation governing if and how such work can be conducted, and this becomes more complex still if special kinds of data are required in the datasets for linkage. Planned work must comply with a number of laws and gain ethical approval at many levels, as well as obtain support from both patients and the public. In order to undertake and co-ordinate such a complex task, there must be a central administrative hub to take responsibility for final study design, applications for ethical approval and requests for data.

In this instance, the project group was based within a health research department, itself within a university environment, providing the study with an academic medical footing. This is not always necessary when seeking to gain access to linked health data, but the credentials of the research group within the academic medical field provided a clear indication that they had the appropriate knowledge and skill set to conduct the planned

research, and therefore optimise the utility of the linkage, which is investigated by the ethical board as an important activity from their standpoint. For them it is important that the linkage is not wasted and its utility maximised as much as possible.

Given the cross-disciplinary nature of the work, and to ensure that outputs were indeed maximised beyond the health sphere alone, a steering committee of key stakeholders was assembled from the outset. This committee was formed to assist in aligning the research objectives with potential wider organisational goals and questions, to provide specialist input into project design and ongoing development, and to facilitate communication and collaboration between the research groups, funders and stakeholders in road safety.

The group consisted of representatives from Imperial College London (the RTI–AID research group), the RAC Foundation, DfT, Transport for London, LAS, NHS Digital, TARN and the British Red Cross. Details of the linkage plan were discussed from the outset, and data flow agreed by the committee.

Without the required organisations or disciplines being in the room at the right time, albeit sometimes only as a virtual presence, questions requiring responses or discussion involving many of the parties could have become difficult and time-consuming, as individual contact had to be made. Perhaps the most unexpected challenge of all was the arrival of the COVID-19 pandemic. As the majority of the research team were clinicians, this effectively halted the project for a short time, and put additional demands on the team thereafter. Even from a non-clinical perspective however, there were certainly further impacts of the pandemic, with both the core, and also the wider group and other interested stakeholders, moving to remote working practices and having to navigate their own new work demands. Outside of the research team and stakeholders, there was a noticeable reduction in accessibility to administrators in the various different organisations. NHS Digital in particular were under great strain, and were not able to be as responsive as previously. This was compounded by a number of staff changes, with key personnel who had been involved in the initial set-up of the project moving on, and new replacements being unable to join committees – and finding the complex mechanisms of the project difficult to process because of the limitations of written applications or emails.

LEARNING POINTS

- *The multidisciplinary steering committee was pivotal* – although having the main research group based in an academic medical department was useful from a patient engagement and ethics application viewpoint, it was more useful still to have access to the wider steering committee. This was an effective means of garnering opinion and officially ratifying support, which was useful when it came to applying for data or ethical approvals. It was also instrumental in building trust between the organisations and, essentially, creating a network that would support the work getting underway at various organisational levels and stages of the application processes.
- *There should be two separate steering groups: operational and data-oriented* – as the project evolved it became clear that committee members were acting in either an operational or data analytical capacity, although both roles were not always represented by an organisation in any given meeting. The recommendation would therefore be for two separate groups: one would be made up from operational leads to help provide oversight, guide research questions for policy and current real-world problems, and assist with applications for data; and a second one consisting of data analysts from each of the dataset organisations.
- *There needs to be greater access to group expertise* – over time, steering committee meeting attendance varied, meaning that administrative heads and data analysts were not always both present. Moreover, the steering committee was intended to convene only at specific milestones, but this proved not to be sufficiently frequent to lean on them collectively for smaller – but nonetheless important – questions. In retrospect, it might have been useful to set up virtual groups using project management or team collaboration software such as Slack, to facilitate more frequent low-key access as well as to house project resources and references. This would have been useful to mitigate any staff turnover.

6.2 Data selection

The study design called for the linkage of transport data and health data to create a linked dataset describing the entire picture from the collision and point of injury, through to hospital transfer, and ultimately on to hospital treatment and discharge. Consideration had to be given to which data fields were to be requested and why. It was necessary to strike a balance between, on the one hand, obtaining a sufficient number of fields to enable creation of a clear and detailed understanding of what happens to injured road users, as well as adequate fields to increase the chance of linkage, but, on the other hand, simultaneously minimising requests for excessive data. Indiscriminate requests for data are not within the spirit of the law, namely the Data Protection Act (see section 6.3).

The linkage methodology will require certain fields for the purpose of maximising linkage (e.g. date of birth, postcode), whilst needing other fields for adequate analysis (e.g. injury outcomes). Others are needed for both linkage and analysis, such as age and ethnicity. The ethics board and the relevant data controllers for each dataset will also require this

justification for each field before they will approve the project and data release. To reduce the risk of re-identification of individuals following linkage, certain fields could be removed (e.g. date of birth) and replaced (e.g. with age) to reduce identifiability.

LEARNING POINTS

- *Data selection must consider the research question, linkage and privacy risk* – although the choice of which data fields to use was based on the research team's needs to answer the research questions, coupled with the need to reduce the identifiability of data coming into Imperial's secure environment for analysis, there was less discussion concerning the fields which would be optimal for linkage to STATS19 and LAS, because the overall emphasis was to ensure the safe data flow, including minimisation of the amount of data flow, for the wider RTI–AID study. Perhaps a better balance could have been struck here with an earlier open discussion about linkage strategy.
- *Greater access to the stakeholder data analyst group would be beneficial* – as mentioned in section 6.1, ensuring opportunities for more focused technical discussion about how to best use the data would have better informed the selection and justification of data fields from various providers, optimised research outputs, and contributed to an understanding of how linkage could be best performed and assessed.
- *Selection of more data fields has cost implications* – a further consideration was data costs, which varied considerably by institution, and were not always clear at the outset. This led to further discussions about which fields and time periods might best serve the research.

6.3 Compliance with legislation

It may be tempting for researchers to request as many data fields as possible, but both the law and cost considerations mean that they will encounter limits. Although cost difficulties can potentially be overcome, compliance with the relevant legislation requires protection of the data subject's privacy by minimising the amount of identifiable (or potentially identifiable) data that can be accessed and processed without their explicit consent.

There are three important laws for consideration in the linkage of health data: The General Data Protection Regulation (GDPR), The Data Protection Act 2018 (DPA) and section 251 of the National Health Service Act 2006 and its current Regulations, the Health Service (Control of Patient Information) Regulations 2002.

The relevance of each is outlined below, together with how RTI–AID study design complied with their requirements. More detailed information on this legislation is also provided in Appendix B.

(i) The General Data Protection Regulation

GDPR came into effect in the European Union in May 2018 and outlined how personal data can be processed in the region. In linkage of health datasets, data can often contain personal data owing to the presence of individual identifiers such as date of birth or address, which

frequently underpin linkage methods. Since leaving the European Union in 2021, the United Kingdom has still retained the same legislation in the form of UK GDPR, although the UK's departure allows the government to independently keep UK GDPR regulation under review.[75]

What does GDPR say?

Article 6(1) of GDPR states that there must be a legal basis for the processing of personal data and Article 9 also sets out the conditions for the processing of special data, including health data. Interestingly, GDPR applies to living subjects only, therefore information relating to the deceased is not considered personal data and is not subject to GDPR.[76]

How does this relate to this linkage work?

Analysis on fully anonymised datasets does not fall under the remit of GDPR as the data is not considered personal. However, in the case of this linkage, there is the need to use identifiers to link records with associated health data, for which there needs to be a defined legal basis.

The ethical and legal ideal is that data subjects should have given consent beforehand, but in the case of large datasets collected for other purposes, this will not be the case – and retrospective consent is not feasible. In view of this, Article 6(1)(e) was invoked for linkage with respect to personal data: the work is to be performed in the public interest. Healthcare linkage and research most commonly falls under Article 9(j), the research exemption, but in the case of this work the potential for public health findings meant that the combination of conditions from Article 9(i) (public health) and 9(j) (research) were used as the legal basis for processing identifiable health information (special category personal data).

(ii) The Data Protection Act 2018

The DPA incorporates additional requirements and safeguards into UK law beyond GDPR.

What does the DPA 2018 say?

With respect to special categories of data, if using conditions (b), (h), (i), or (j), as is the case here, the work must also meet Part 1 of Schedule 1 of the DPA 2018.

This means that to safeguard individuals in the datasets, research must be conducted on minimised data (i.e. data which contains no more information than that which is directly relevant and necessary for analysis to answer the research question). Additional safeguarding should also include pseudonymisation to reduce the possibility of re-identification of individuals during processing.

How does this relate to this linkage work?

To comply with DPA 2018, as previously outlined, the request for each data field had to be justified by stating how it was required for either linkage or the later research analysis. Where a field was required for linkage alone, it was to be removed from the dataset after completion of linkage, and the linked dataset then pseudonymised prior to the research work.

(iii) Section 251 of the National Health Service Act 2006 and its current Regulations, the Health Service (Control of Patient Information) Regulations 2002

Health data with potential identifiers not only counts as special category data under GDPR but also falls under the definition of confidential patient information as defined by section 251 of the National Health Service Act 2006, which applies to both the living and deceased.[77]

What do the National Health Service Act 2006 and the Health Service Regulations 2002 say?

Under the common law duty of confidentiality (built up from case law through individual judgements) when information is given in circumstances where it is expected that a duty of confidentiality applies, which is the case for health records and submissions to health datasets such as HES and TARN, information cannot usually be disclosed without patient consent.

This requirement can be waived where confidential patient information is to be used to benefit research, provided there is section 251 support. This support permits processing for a range of medical purposes, broadly defined to include preventative medicine, medical diagnosis, medical research, the provision of care and treatment and the management of health and adult social care services.

Section 251 recognises that there are essential health service activities, such as research, that require the use of confidential patient information when neither anonymisation nor retrospective consent are possible. In such cases section 251 permits the temporary lifting of the common law duty of confidentiality for medical purposes.[64]

Section 251 support is obtained through application to CAG, which acts under HRA. CAG reviews applications for research and non-research work, and determines whether there is sufficient public interest to temporarily lift the duty of confidentiality.

How does this relate to this linkage work?

The data requested from STATS19 and LAS did not contain personal identifiers, nor was it considered to be confidential patient information. However, in its raw form HES and TARN data were considered as confidential patient information and needed to remain defined as such for the planned linkage. NHS Digital was the data controller for HES and regularly performed HES–TARN linkage for audit purposes, as well as conducting periodic HES–STATS19 linkage. It also offers bespoke linkages on application. Therefore, to capitalise on this prior work and reduce the risk of breaches in confidentiality, STATS19 and LAS data was to be sent to NHS Digital so that they could perform the linkage. This would require section 251 support, but once the linkage was complete the linked dataset was to be pseudonymised and sent to the research team at Imperial College London.

It should be noted that under GDPR, pseudonymised data is still considered to be personal data, as there is the potential for re-identification with the pseudonymisation key. However, to effectively remove this possibility the key is stored separately, with NHS Digital, without the research team having access.

LEARNING POINTS

- *Multilateral design with stakeholders including data controllers is constructive* – the very first steering committee meeting discussed the legal requirements for this putative linkage but was also able to build upon a tentative data flow put forward by the research group with various comments from the stakeholders present.
- *The inclusion of representatives and linkage leads from NHS Digital is essential* – the riskiest component of this linkage is the processing of special category health data and confidential patient information required to make the linkage. To minimise this risk, the data flow design hinged on NHS Digital receiving the datasets from DfT, LAS and TARN, and thereafter performing the linkage itself, and providing only a pseudonymised dataset to researchers. From early steering committee meetings, NHS Digital voiced this as their preferred data flow and stated that once ethical approvals were obtained, this could be done.

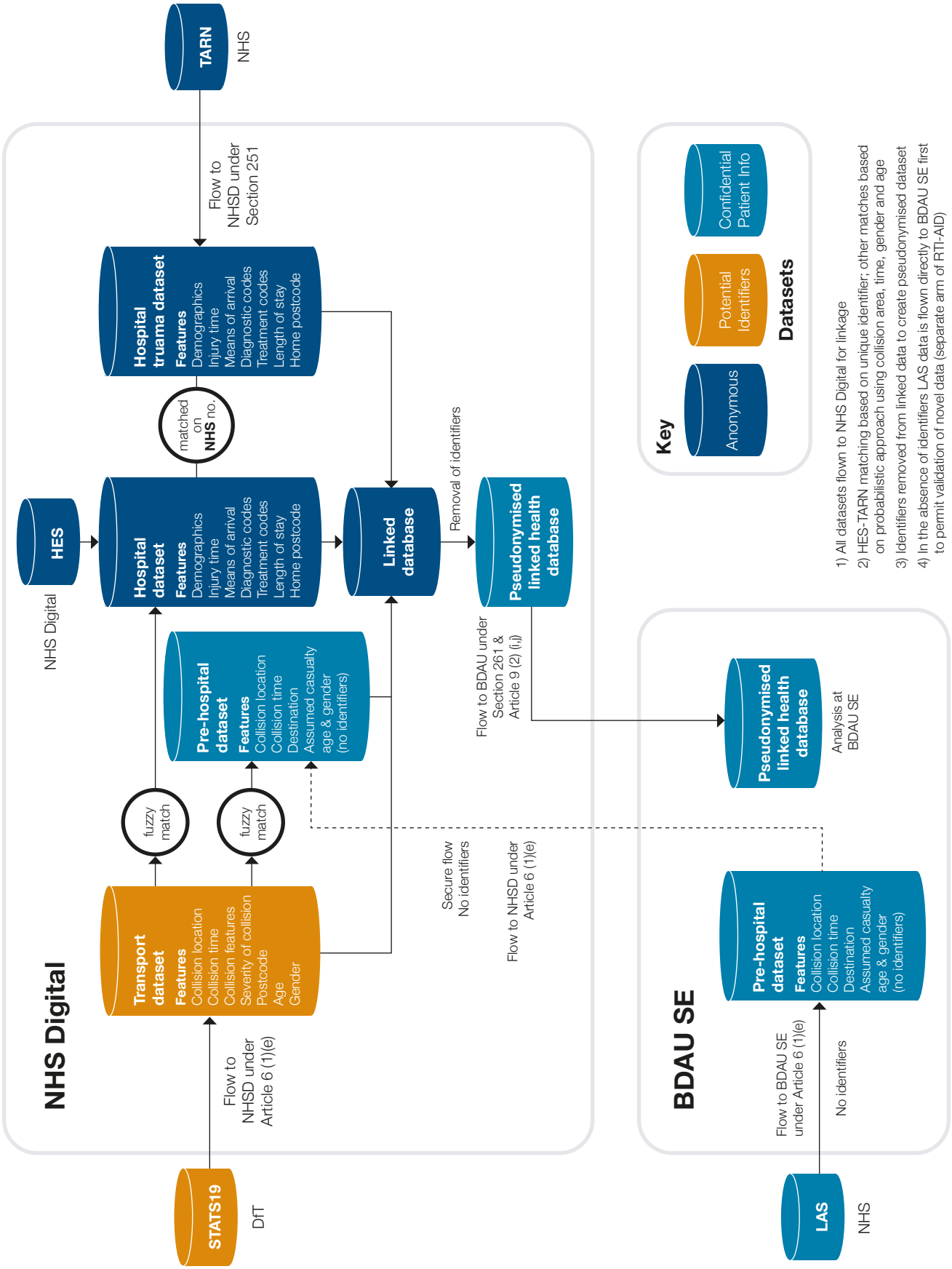
6.4 Data flow

The design of an optimal data flow must consider which data must travel from their original controller at the parent institution, to a secure destination for either linkage or research. It must also consider the legal basis for each flow and process, as well as the storage conditions for data to be processed and housed.

In this instance, bearing in mind the legal requirements, and the analysis required, a data flow had to be designed to allow both the linkage to take place securely by NHS Digital, and the analysis to be conducted by the IGHI team. Figure 6.2 shows this data flow, and all the datasets showing their status as concerns identifiability.

Because the RTI–AID project had a dual objective of creating a gold standard linked dataset from transport and health institutions, and validating novel crowdsourced digital data for road safety, two parallel data flows were created and submitted together for ethics approval. Here, for clarity, only the data flow for the provision of the linked gold standard dataset comprised of STATS19, LAS, HES and TARN is described.

Figure 6.2: RTI–AID data flow for linkage of STATS19, LAS, HES & TARN



Source: Author's own

STATS19 with postcode to optimise linkage was to flow directly from DfT to NHS digital owing to the presence of this identifier. LAS data without identifiers but containing information on collision features, location, times and potential severity were provided by LAS to The Big Data and Analytical Unit Secure Environment (BDAU SE) for analysis against the novel crowdsourced data, as well as supply to NHS Digital for linkage to HES. TARN data was to go directly from its data controllers to NHS Digital for linkage with HES, because it contained confidential patient information.

Upon receipt of STATS19 and LAS data, from DfT and Imperial respectively, NHS Digital was to link this to HES using fuzzy matching (i.e. probabilistic methods). HES and TARN data were to be linked by NHS number, as has previously been performed, and the entire linked road safety dataset of STATS19–LAS–HES–TARN was to then be pseudonymised. The pseudonymised linked road safety database would then flow from NHS Digital to the BDAU SE at Imperial College for analysis of how particular collision and demographic features resulted in different health outcomes.

The BDAU SE is ISO 27001 certified and compliant with NHS Digital's Data Security and Protection Toolkit, providing a platform for safe data storage and processing. All users are also required comply with annual information governance training and re-certification. In line with GDPR and health research ethics, the linked data would be kept only for the period of analysis and publication of findings. Thereafter it would be archived for a limited period in line with the study period and finally destroyed.

LEARNING POINTS

- *Interpretation of legislation remains complex* – despite iterative multi-stakeholder co-design of the data flow, and subsequent ethical approvals of the study design and data flow by HRA and CAG, at the final stages, neither NHS Digital nor DfT could confirm the legal basis for the flow of data with identifiers from DfT to NHS Digital. This flow had previously been executed for prior STATS19–HES linkage, but the new NHS Digital team was unable to agree on whether this same basis could be used in this new context. As this component of the linkage did not include confidential patient information, CAG was unable to comment, leaving the data flow and linkage on hiatus.

6.5 Privacy and the public

For there to be ethical approval to possibly use health data, and in particular confidential patient information, the HRA and its Confidentiality Advisory Group (CAG) must be convinced that the benefits of the research outweigh the risk of taking steps involving confidential patient information without consent.

Although researchers can outline these benefits, and the input of interested parties – including healthcare providers, government agencies and public health practitioners – can support the arguments, no voice is more important than that of the patient and the public. There needs to be evidence of testing the acceptability to patients and the public more

generally of using confidential patient information without specific consent for the purpose of this research.[46,47,77]

Patient and public involvement is an example of the use of lay people in shaping and influencing research. Whilst the public were not involved in this way for RTI–AID, a number of engagement activities took place to disseminate the aims and methods of the work and seek public comment.

6.5.1 IGHI Patient and Public Involvement and Engagement (PPIE) Group

IGHI has a group of members of the public, including previous patients, who actively wish to be engaged in upcoming medical research projects at Imperial. The study proposal, including its aims and methods, together with information governance considerations, was sent out to this group for comment and feedback. The proposal received written positive support from this group.

6.5.2 PPIE Café

The IGHI PPIE team also organised a half-day open event showcasing some of the upcoming research at a public space in central London. This PPIE café took place with lay helpers assisting the researchers to outline and explain their research for walk-in members of the public to comment on and ask questions. Again, the work received positive support. There was also an interest in the public being given the chance to put further questions about the linked dataset in the future.

6.5.3 RTI–AID web page

A dedicated RTI–AID webpage was set up as part of the IGHI BDAU pages on the Imperial College website. This outlined the project aims and mechanisms, as well as setting out how the data would be used. Additionally, it set out information on how people could ask for further details, or object to the work.

It is an important requirement for HRA and CAG approval that the public have the ability to opt out, should they wish to. The RTI–AID website provided details for the research team to be alerted, but in addition each of the individual NHS organisations (London Ambulance, NHS Digital and TARN) all have pages on how the public can request for their data to be removed or not used for certain purposes.

LEARNING POINTS

- *Dedicated in-person PPIE activities can overcome study complexities* – explaining the complexity of this project, specifically regarding the linkage, the use of confidential patient information, and re-identification, was challenging but necessary. Establishing a lay person's understanding of the data flow and risks related to re-identification needed to be confirmed before inviting comment regarding the public benefit of the research. For the lay audience, this explanation was best done verbally, rather than via provision of written material. The PPIE open half-day was therefore instrumental in obtaining a diverse range of public responses. Even then, in some cases, understanding was not complete and more patient- and public-facing reference materials could have been useful.

6.6 Ethical approval and data permissions

Once the study design and research proposal were complete, they were formally submitted for ethical approval by HRA. The application had to include the rationale for the work, a list of those contributing to the work (including co-ordination with the groups in the steering committee), the data being used, the reason for its use, a clear data flow with detailed explanation of the processing, and planned outputs/benefits. At this stage HRA provided only an interim approval, as section 251 support is also required.

Thereafter, a follow-up application was also made to CAG. This had to include evidence of PPIE activity with public support, the means for public dissent or opt-out, and a letter of support from the Caldicott Guardian (a senior person responsible for safeguarding the confidentiality of healthcare data) for Imperial College London. The CAG application also included a brief interview process in which CAG representatives ask about the work, the justification for requested data, plans for analysis and its implications for the future, as well as raising any concerns they have. RTI–AID was granted final HRA and CAG approval with section 251 support in 2020.

Although not required for the ethics process, as part of good practice and to comply with GDPR and the DPA, a Data Protection Impact Assessment (DPIA), also known as a privacy impact assessment, was also carried out. The DPIA is a risk assessment that helps to identify and minimise risks relating to personal data.

Finally, once HRA approval with section 251 support was given, data could be requested from the relevant custodians. Each organisation has its own separate application for data requests, as well as ethical approvals, although provision of HRA and CAG reference are the most useful and important information to provide, as well as the overall data flow.

The presence of representatives from relevant organisations on the steering committee was helpful in making applications, but nonetheless individual applications still had to be made and data agreements formalised. The most complex application was to NHS Digital via the Data Access Request Service (DARS), not least because the application was needed to outline the

plan for bespoke request for linkage.[78] The information provided to DARS was expected to demonstrate the legal basis for access to the requested data: that researchers would comply with information governance requirements; that the data would be securely stored; and that it would be use to improve health services and not solely commercial purposes.

LEARNING POINTS

- *The revolving doors for ethics approval from CAG and HRA are problematic* – the requisite ethics process is onerous, involving extensive application forms, multiple steps and in-person appointments to secure approval. Furthermore, the process is not linear, and applicants frequently find themselves needing to go back and forth between steps involved in building the plan and gaining interim approvals along the way – for instance, initial preliminary HRA approval is required, followed by CAG section 251 support before final HRA approval. This means that CAG approval cannot be sought without an initial HRA application, but final HRA approval is not given without CAG. Another example is ethics approval, which is not possible without evidence of PPIE support; but at PPIE events the question will be posed as to whether ethical approval has been gained – another catch-22 situation.
- *Steering committee members or stakeholders with ethics experience for linkage are invaluable* – establishing a tentative data flow, following steering committee meetings with NHS Digital and other data controllers present, meant that the legal and ethical considerations of each component of the data flow were considered early and by key figures at each organisation. This saved further repeat steps at later stages.
- *There is benefit in seeking to get as many steps as possible in the process agreed, ahead of the ethics submission* – this includes those described here: data selection, legal considerations, data flow, PPIE and the record linkage plan, as well as Caldicott Guardian support. At every turn, it can feel as if one cannot proceed without having all other steps complete which can create a sense of paralysis. The principal finding in this respect is to at least start and expect some degree of iteration, although having early wide stakeholder discussions and commencing a detailed project map is most useful.

6.7 Record linkage

Perhaps one of the greater weaknesses of the RTI–AID process was the limited nature of the discussion of the mechanism for record matching and data linkage. This is not to say that the research group had not given consideration to how the different datasets might be problematic, but early on there was a recognition that the linkage would be performed by NHS Digital's in-house team, which performs bespoke linkages, building previous relevant linkages they had performed using STATS19, ambulance data, and TARN.

Previous sections of this report and Appendix A outline how the datasets differ in terms of the population they capture. It was expected that not all STATS19 cases would be linked to health data, and that if anything it would predominantly be the most seriously injured and killed that would feature. Whilst this is an important group, equally worthy of study are the

individuals with more minor injuries, who are more likely to be missing from one or two of the datasets, if not all of them.

To appraise this fully, ideally the research team would have been able to have more direct involvement in the linkage strategy and its iterations, enabling them to follow how and why cases are matched or missed; but this was not an explicit part of the final study design.

LEARNING POINTS

- *A closed-door approach to record matching by NHS Digital would probably have undermined the potential for optimal linkage* – although linkage is rarely perfect, the focus on creating a low-risk data flow with no confidential patient information outside of NHS Digital overshadowed discussion about how to create the best linkage. Although it is possible that further discussions would have taken place between the research group, the steering committee and NHS Digital once the linkage was underway, this was not specifically planned.
- *Unmatched records are still important for analysis* – inevitably not all records will be matched, even with optimised rates of linkage. However, unmatched records still need to be assessed and, where possible, included in analysis as they provide an important aspect of the broader road injury picture. These are more likely to be more minor injuries or delayed presentations, but the burden of injury for this group has to be appreciated if efforts to encourage active transport through cycling and walking are to be maximally effective.[79,80]

6.8 Summary of RTI–AID linkage work findings

Chapter 6 has described, step by step, the extensive process required for linkage of the existing range of official road safety datasets from police, ambulance and healthcare to comprehensively describe the road safety situation in the UK. Nonetheless it is only a summary, and conceals the many further discussions, debates and iterations required to achieve success at each stage.

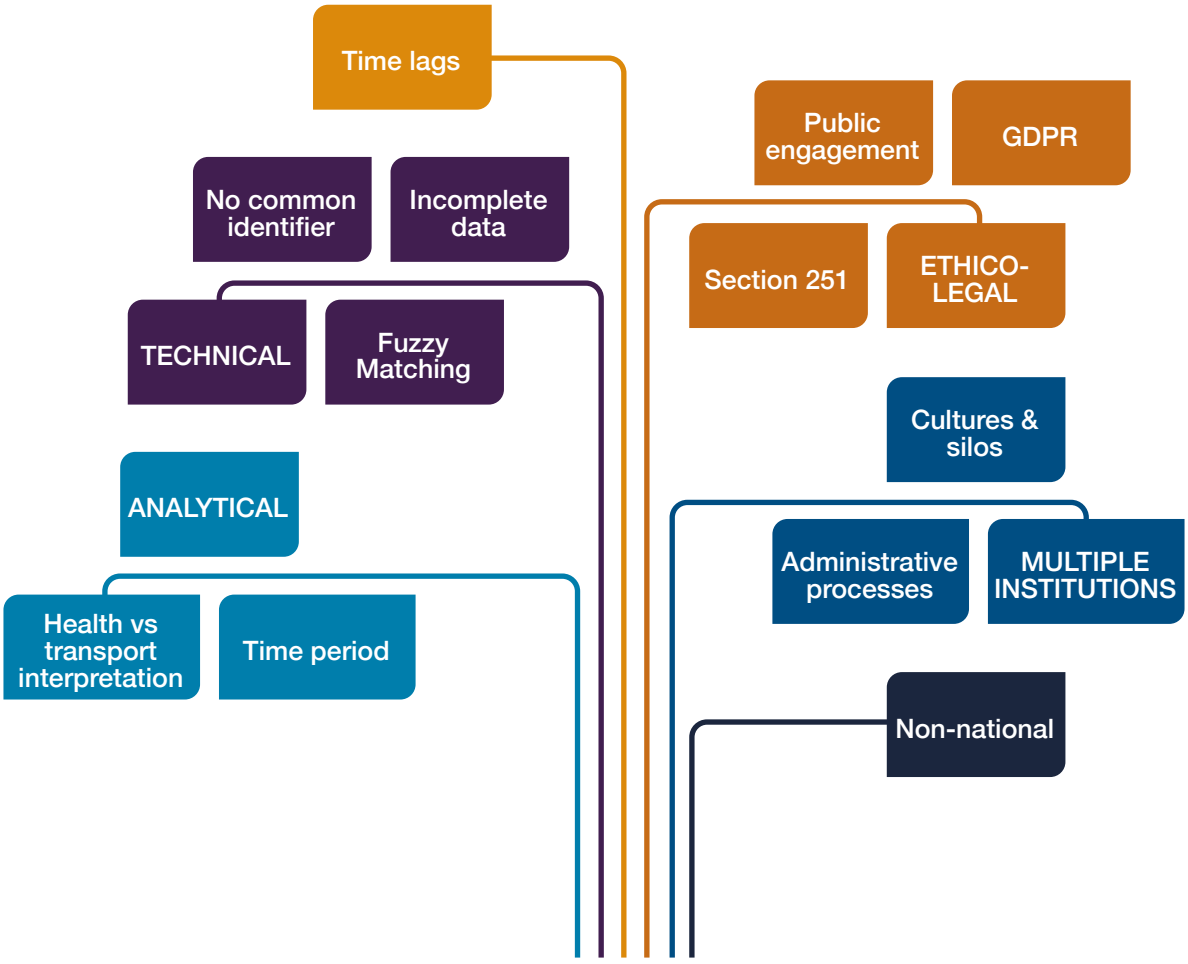
As a linkage this extensive had not been attempted before, either within the research organisation or in the wider road safety field, there were few people or resources to refer to when mapping out the process and establishing how it might vary for the specific work being conducted. The project began in 2018, at a time where there was palpably less guidance and encouragement about linking large governmental and health datasets, as well as an appreciable general anxiety about the risks involved. In this case, the draft application to HRA was planned very soon after the first steering committee meeting, before the research team had gelled with the wider stakeholder groups, and conversations about how to go about the work had matured. Also, section 251 support for research in the department had not been sought, and consequently, despite the presence of a large steering committee and an experienced research team, this was still new territory.

To build on this work with further attempts at linkage in UK road safety, there must be a clear understanding of the relevant regulations and legislation concerning the data flows and

processing. GDPR is complex, and even with the expertise available amongst the research organisation, the steering group and the ethical board, there were at times differences of opinion – even between individuals in the same organisation – on how it would apply and which articles were to be used. There must be adequate expertise and consensus to confirm all aspects of the plan by the time it is presented to ethics boards and organisations for data release or processing.

The complexity of the ethico-legal mechanisms for gaining permission for health data linkage is multiplied when this linkage is required across more than two datasets, as in this work. This complexity arises from the types of data; the legal basis for its processing, including both research analysis and linkage; and the need for identifiability in some datasets to optimise linkage. However, this complexity does not mean that the work should not be undertaken. Indeed, the public appetite for it appears to be there, and the ethical approval and section 251 support gained for RTI–AID demonstrates that there is a means to negotiate the important linkage between STATS19, ambulance data and hospital data in the form of HES and TARN. This should be built upon to create a longer-term plan for a comprehensively linked road safety dataset which could benefit researchers, policymakers and, of course, the public.

Figure 6.3: Barriers to linkage across datasets pertaining to road safety



Source: Author's own

7. Galvanising the UK Road Safety Community for Definitive Data Linkage

7.1 Collective action must lie at the root of a paradigm shift

There needs to be a radical change in our expectations of road safety data in the UK, and this needs to happen now. Today, our goals, and those of our neighbours and economic counterparts further afield, centre around bringing fatality and serious injury on the roads down to zero. This is no small feat. Data is the foundation undergirding how we assess our progress, target interventions, and measure their efficacy in this campaign to be rid of serious traffic injury – but considering what the road safety community can currently access, how much can we truly know? Road safety advocates, researchers, policymakers and practitioners must come together to push for the establishment of the kind of comprehensive linked data that other groups in government, public health and industry are already exploiting.

Currently accessible data may accurately establish fatality rates, which are clear endpoints, but if each fatality is associated with a further number of seriously injured people, how much is really known about the injured cohort? Assessment of MAIS3+ road casualties is not a trivial calculation, and a lay police officer cannot be expected to reliably make this clinical judgement. An apparently innocuous minor head injury or concealed pelvic fracture may not immediately come across as serious, but either could develop into life-threatening presentations with time. Consequently, the most viable next step to understand the injured is to comprehensively link STATS19 to ambulance and hospital datasets, each with its own part of the puzzle to report on. A comprehensive definitive linkage which compiles STATS19, pre-hospital ambulance data, HES and TARN is surely the clear target to be aimed for, because without including each of these datasets, we compromise our understanding of the breadth or depth of the problem, and potentially even reduce rates of linkage.

The road safety community is at an impasse, with reams of useful data and powerful tools for its analysis visible ahead, but a convoluted ethico-legal process for its access firmly planted between where we are now and a fruitful exploitation of that data. In recent years, public concerns about trust and privacy have generated a further minefield, deterring many researchers from attempting what seems like a logical objective of linking transport and health data. The RTI–AID research group was made up of health practitioners based at a leading academic institution with a strong track record in the analysis of big data and a solid working relationship with NHS Digital, but, equally, this undertaking was of massive proportions, and no other group at the institution had attempted such an extensive linkage before. In the end the planned linkage fell just short of fruition, but the project has been able to establish an important roadmap for future efforts.

Although the role of health professionals in the study group and a base at an academic institution, in fact the home of a leading trauma centre, did ostensibly confer some advantage when it came to obtaining ethical approvals, the greatest learning point from this work has been the critical nature of a broad multidisciplinary steering committee. The committee's strength came from its breadth, composed as it was of a range of stakeholders with a broad understanding of road safety, as well as operational and data leads from organisations controlling data. Each member had invaluable institutional knowledge relevant to different parts of the process, but together this group was able to sense-check plans, intelligently discuss issues, and offer solutions to even minor but nevertheless important problems that could have derailed linkage success. The intended data linkage was a manifestation of co-operation between these various groups. In the end, progress was halted by a failure to reach consensus on a minor aspect of the data flow, at a time when the committee could no longer readily convene owing to pandemic-related logistics and availability.

Whilst reassembling a similarly relevant national committee would be the ideal means of continuing progress toward the establishment of a definitive road safety linkage, it is equally important that the wider road safety community too echoes calls for a comprehensive linked dataset. The community will need to speak as one voice to bring about a broader movement demanding high-quality data that is not limited to a small group of academics. To date, road safety researchers have attempted smaller-scale linkages with varying success, but this

has involved a laborious duplication of effort, with each group applying for data access with degrees of linkage that vary owing to the limited presence of identifiers. This heavy lifting can and should be performed by data experts with the appropriate skill set and access to identifiers, namely those dedicated to working on health data linkage at NHS Digital. Coupled with ongoing dialogue with the steering committee regarding the design and assessment of the matching process, use of NHS Digital linkage expertise is the best route forward for establishing a national gold standard linked dataset. Furthermore, once constructed, the government recommendation is for these rich pseudonymised datasets to be made accessible to a wider range of researchers through secure or trusted research environments (TREs).[81] In this way the country's road safety researchers can focus their efforts on targeting the stubborn plateau in road fatalities rather than facing a quagmire of bureaucracy.

TREs could also bring a whole new tranche of healthcare researchers to the road safety field, providing a different perspective on the problem. For the health researcher, new but important questions arise: are the right road casualties being triaged to the right centres? Why do some demographic groups fare worse than others following similar collision circumstances? What medical interventions can be introduced to improve clinical outcomes for particular injury types? This additional contribution would be most welcome in tackling remaining levels of serious road traffic injuries. Moreover, TREs would also provide the opportunity for emerging analytical tools, including machine-learning methods which could provide additional insights that we have not been able to appreciate to date. One particular application which the RTI–AID group had hoped to apply to linked data is the development of scoring systems that could be used by police or paramedics at the roadside to predict possible injury severity on the basis of collision factors and basic immediate clinical assessment. Such a development would be an exciting departure from looking at population behaviours and preventative strategies, and would further open up the field of emergency post-crash care in real time in a way that is informed by intelligent use of data.

All these opportunities, with their huge potential for making great advances, presently lie before us. The RTI–AID project has demonstrated that there is an administrative, legal and ethical means of conducting a useful road safety data linkage with public support. Limited resources halted further pursuit of the linkage by the research team in this instance, but with road collisions costing £33 billion per year (1.5% of GDP), predominantly through human cost,[82] there is an obvious financial incentive for the country to reattempt this endeavour. Disregarding or forsaking a collective drive for data linkage in the hope that old methodologies will be able to offer as much innovation is a mistake. Data does indeed save lives. It therefore stands to reason that efforts to reduce road deaths and injuries should start with a demand for better data.

7.2 How to exploit the opportunities of today

The organisation and processes required to achieve a nationally linked dataset may seem daunting, but there is much in favour of striving to reach this goal. This report outlines the steps that have shown it to be possible at a regional level, but there are further factors that should encourage the pursuit of this linkage by all those concerned.

7.2.1 Governmental appetite

A range of governmental reports and guidance published over the past five years has encouraged the creation and use of linked data, particularly of government and health datasets, to address some of the societal problems that can be explored in this way. The mood appears to have transitioned to one of active support.[18–21] The Digital Economy Act 2017 facilitated greater data sharing across government, but did not extend to the NHS. However, there has been general acknowledgement that there has been less progress when it comes to a wider linkage of health data. Indeed, research is being conducted to find out what the remaining barriers to linkage might be. These have highlighted ongoing departmental anxieties regarding whether data can or should be released, and concerns about what happens in the event of problems; but this is to be countered by robust planning with comprehensive ethico-legal considerations and clear public engagement for the specified linkage.[81]

7.2.2 Public expectations, privacy and trust

Most studies demonstrate widespread but conditional support from the public for data sharing for linkage in health datasets. There naturally needs to be clear, transparent explanations of the rationale for linkage, the mechanisms involved, and a risk analysis for the sake of the public with any linkage project using health data. However, in this information era, as data volumes increase and findings from unstructured data demonstrate how it can be used in interesting ways, there is a greater expectation from the public that data held about them should be used for the public good. It is no longer acceptable to keep such information stored away. Certainly privacy and confidentiality have to be upheld, but there are ways of doing this that still allow for the analysis of public data in a way which benefits the people to whom it belongs.[83]

Following the suspended 2021 General Practice Data for Planning and Research (GPDPR) programme, research suggested that the public not only recognised the value of their data but also wanted to make it available for use in research. However, what they clearly objected to was the way in which GPDPR had not permitted them to decide on data opt-out, or allowed them to have their say on with whom their data would be shared and for what purpose.[84]

The Government thus commissioned Professor Ben Goldacre to investigate how to safely balance the unlocking of the potential benefits of research on health record data with the need for patient confidentiality.

The Goldacre review, *Better, Broader, Safer: Using health data for research and analysis*, recommended the use of 'trusted research environments' (TREs), which are highly secure computing environments which allow approved researchers to conduct their analysis remotely.[85] TREs, which have been used widely in other sectors also, reduce the need for multiple applications and data releases and go further in mitigating patient and public concerns regarding confidentiality.[86] Last year the Department of Health and Social Care launched a new strategy to regain public favour and trust. *Data Saves Lives: Reshaping health and social care with data* is intended to allay residual fears from GPDPR – but with ongoing public misgivings about how the Government may be contracting external companies to run data platforms, its success remains to be seen.[87]

That said, the linkage of road safety data is a very clear and distinct use case, with the linkage being almost entirely for the public benefit and with little likelihood of the data being useful for commercial purposes. RTI–AID PPIE activities were successful, but could be greatly expanded to create a groundswell of patient and public support, especially in view of the high volumes of younger people affected and the range of road safety advocacy groups that are well placed to take part in engagement activities. Ideally, with sufficient time and planning, patients and the public could also make contributions to research questions. This would constitute true public involvement in research. This is an ethical ideal, and should be planned for where feasible.

7.2.3 International competition

Calls for increased data linkage are not limited to the UK. In 2002 the World Medical Association set out its 'Declaration on Ethical Considerations regarding Health Databases and Biobanks', encouraging the sharing of health data for 'secondary' uses.[88] Two decades later the International Population Data Linkage Network facilitates communication and the exchange of ideas between centres specialising in linkage, with its members hailing from a range of countries across the economic spectrum.[89] Best represented are the United Kingdom (particularly in the health field), Australia, New Zealand and Canada.[89,90] This represents a significant overlap with countries with a strong track record in road safety, and also evidences an emerging practice of data linkage of transport and health data for this purpose. In some cases there is an established nationalised linked database, such as Sweden's STRADA.[16] In other cases researchers and stakeholders have been truly innovative and linked various hitherto unrelated databases to make national recommendations, such as New Zealand's SORTED (Study of Road Trauma Evidence and Data) study with seven databases incorporated.[17] Even the United States, chronically behind in road safety performance when compared to European or Australasian counterparts, has begun evaluating the linkage of police, ambulance and trauma registry data.[18]

The STATS19–HES linkage set up in 2012 was a landmark achievement but since then the UK has fallen behind. In the intervening decade little progress has been made, and although that linkage is still used, it is primarily to verify the quality of STATS19 injury coding, rather than for the purpose of understanding how the nuances of road dangers are unfolding.

7.2.4 Funding

Hand in hand with the growth of national and international appetites for linkage, significant associated funding has been raised to support groups attempting this work. In the UK, Administrative Data Research UK (ADR UK) funds multi-stakeholder linkage projects and is in turn funded by the UK government.[91] Its emphasis is on releasing the knowledge locked in public sector databases, which could make cross-cutting research possible through linkage. Output datasets need to inform policy and have the potential to have a significant positive impact on lives. The stated exemplars of public sector data linkage in Chapter 2 were all funded by ADR UK, and the proposed comprehensive linkage for road safety datasets fits well with what ADR UK and most funders supporting linkage are seeking to achieve.

7.2.5 Technical innovation

Finally, recent years have seen a dramatic rise in the affordability of computer processing power and the availability of more advanced linkage techniques. Until the end of the twentieth century, Fellegi & Sunter's algorithm for probabilistic linkage (1969) was the most widely used, but the rate of record linkage research has substantially increased since then.[27] This has been driven in part by the size of available datasets. The use of neural networks and higher-level matching features can substantially improve matching in certain circumstances. In addition to matching, statistical and machine-learning techniques may also permit greater use of the rendered linked dataset, allowing researchers to get the most out of it[92,93] – for example, which collision features lead to which injury types and severities, thus allowing police or paramedics to forewarn hospital doctors about a particular casualty's potential to be significantly injured or deteriorate.

The RTI–AID project sought to examine how unstructured crowdsourced data compared to a linked standard dataset, examining road collision and injury outcomes from a combination of transport and health organisations. It also intended to examine how the linked dataset could reveal new information about cause and effect in road collisions and injuries, as well as the influence of pre-crash and post-crash factors in different demographic groups. Although resource constraints precluded an analysis of the linked data, the project's efforts to link STATS19, LAS, HES and TARN datasets demonstrate that this linkage is possible, albeit with considerable administrative efforts needed to obtain the correct permissions and approvals. Regardless, this experience is invaluable and the lessons learnt from this pilot must be built upon to take things forward in a multidisciplinary, multisectoral approach, establishing a new comprehensive linkage that can serve road safety researchers, policymakers and the public in the years to come.

7.3 Recommendations for a shared way forward

The RTI–AID project provided a foundation on which to build, but the next stage needs to be a transition from the exploratory phase to one of pre-planned strategy. In addition to the experiences and learning points from the main project highlighted throughout Chapter 6, specific recommendations the for next steps are described here.

7.3.1 Multisectoral, multidisciplinary stakeholder steering group

The steering group was pivotal in ratifying the agreement, organisation and justification for the linked STATS19–LAS–HES–TARN dataset. Although a unilateral group could take on an attempt to organise such a linkage, without the awareness and support of and interaction with the relevant parties, this would have been a challenge when it came to obtaining ethical approval, public support and receipt of datasets.

If a concerted effort to formalise this linkage for wider use is to be made, there must be a steering group of similar stakeholders from the relevant data-owning organisations, researchers, policymakers and people on the ground in transport and health, as well as other proponents of road safety, including non-governmental organisations and public or patient groups.

Similarly, health professionals, ranging from public health to clinicians, are needed to more actively engage with this work, since much of the ethical approval hinges on demonstrating how there is public health benefit, and in this matter both sub-disciplines have contributions to make. In conjunction with this, health data linkage expertise is also useful, especially as linkage within this field is some of the most complex and dates back the furthest.

This group needs to clearly state the mutually beneficial nature of linkage, and how it benefits the road safety community and therefore the public. Further to this they need to also be able to advise on how the data flow may work, as well as which data fields are to be considered for matching and analysis, and which should be removed ahead of pseudonymisation. As a cohort they must agree on how to strike the best balance between data availability and data security in a way that considers the public interest. What this balance is needs to be debated, and may change over time.[94]

Whilst the wider group could meet at key stages in project design and development, between such meetings ongoing multilateral conversation should be facilitated, perhaps through group messaging using work platforms such as Slack. This access to different groups and disciplines would vastly speed up the process. Furthermore, subgroups could be created for the various organisations or disciplinary groups such as data analysts or operational leaders.

Although dedicated researchers will ultimately need to make the required administrative applications, the steering committee needs be at the heart of decision-making. One of the prevailing barriers to linkage at present is the often-siloed nature of the working practices of the different groups in road safety. Where collaboration does occur, it is at regional level, for a limited question or a single linkage. The organisations who own these datasets need to come together to push for linkage of all four national datasets as a long-term collaboration. This can then be made available to researchers and policymakers to enable them to more aggressively tackle ongoing problems associated with reducing road injuries and fatalities.

7.3.2 Co-design of linkage strategies

The work described here deferred much of the linkage to NHS Digital; however, for establishing a national linked road safety dataset, there would need to be a very clear discussion and open forum with respect to the data fields used and the linkage methodology. With RTI–AID, identifiers from DfT and LAS were limited to permit work aside from the main linkage; however, the presence of identifiers such as home postcode have been important in increasing linkage rates in other studies. As a national linkage strategy would be focused purely on safe linkage of the national datasets, useful identifiers could be sent directly from data controllers to NHS Digital to perform the linkage.

Given the importance of finding the right thresholds for confirming a linkage or assuming a match, researchers or data analysts who know the datasets ought to be involved in linkage methodology design, as they will also have to make a judgement call on how the linked data can then be interpreted. This is good practice in the context of dataset linkage for social and health policy. One way to do this might be the provision of pseudonymised ‘toy datasets’ that the analysts could work with to try out different linkage methods and set varying linkage

thresholds. Again, a co-design – whereby the methodology is clearly published and linkage error assessed – would meet the prescribed guidance on how linkage should be conducted. Once the linkage methodology has been agreed, the linkage could then be conducted in-house at NHS Digital prior to creation of a pseudonymised dataset that can be provided externally for analysis.

7.3.3 Legal expertise

GDPR and the related legislation that must be complied with for this linkage are complex. Each step in the process, from data flow between institutions, to linkage, to processing and through to analysis, must be justified and have a legal basis. Sometimes it can be unclear which article applies, and there can even be disagreement between organisations as to which part of legislation is most applicable, if indeed any. Although each institution will have its own data protection officer, it would be useful to appoint an independent legal advisor to confirm, following stakeholder discussion, that the right balance has been struck between planned data sharing, linkage and the risk of re-identification. Moreover, this would both add weight to ethics applications and confirm legality while data flows and processing are taking place. This is particularly useful for data flows from non-health organisations, as both HRA and CAG may feel that their remit does not extend to non-health datasets.

7.3.4 Patient and public involvement

Patient and public engagement is the dissemination and discussion of information about research to the public. Patient and public involvement is where research is conducted in collaboration with the public, who actively influence the research.

Patient and public engagement is adequate to obtain ethical approval and begin the linkage process; however, the aim should ideally be for patient and public involvement.[96] In the context of creating a national STATS19–Ambulance–HES–TARN dataset, this becomes even more important for a number of reasons.

Explaining linkage, especially a complex one where there is the need to use identifiers within confidential patient information, can be difficult to explain to the public in a way that everyone can grasp. How the work is framed is highly influential on how the public feel about it.[95–97]

Patient and public engagement activities should also continue with dedicated webpages at the institutions of the parent datasets, and ideally more public-friendly ways, such as videos, of explaining the project, as have been created elsewhere between research institutions and the Confidentiality Advisory Group (CAG).[97]

7.3.5 An overarching aim for a national dataset

Work to date, including RTI–AID, has focused on regional linkages for a number of reasons, but the next step at this stage must be towards a national linked dataset. Further preliminary work or toy datasets could be generated with a smaller scope, but the longer-term vision must be to make the national dataset available to researchers and policymakers across the country. As the mood within both the government and the NHS slowly warms to linkage, and the public expect more from the data that is held for them, these expectations must be met. A comprehensive multi-stakeholder group effort must not be wasted by limiting linkage

to a small region alone as the this would fail the wider public. A growing number of other nationally linked datasets and calls for greater linkage of public datasets would suggest that this goal is also reasonable.

In the case of linkage to ambulance data, this has previously been restricted by the fact that ambulance data sits with the local ambulance NHS trusts, at a regional level.[98] However, as of 2022, NHS England has created the national Ambulance Data Set.[99] This contains the same data fields used by RTI–AID but now provided by individual ambulance trusts to NHS Digital to create the national dataset. This now means that linkage across the four key road safety datasets, including the ambulance component, is eminently possible at the national level.

Aside from the datasets also being native to each home country, the application process for data access, and in this case linkage, may need to be made on a country-by-country basis. NHS Digital and its DARS data request service are available for England and Wales, but for Scotland and Northern Ireland applications would need to be made to the Electronic Data Research and Innovation Service and the Northern Ireland Statistics and Research Agency respectively. With respect to section 251 support, a similar process is available in Scotland via the Scottish equivalent of CAG: the Public Benefit and Privacy Panel or Health and Social Care.[93]

8. Future Context and Conclusions

There has to be a paradigm shift in how we use governmental transport and NHS datasets relating to road safety. With health outcomes front and centre stage in the challenge presented by road collisions and injuries, it can no longer be acceptable to view this data in isolation. Instead it needs to become part of more widespread ambitions to join up data across the public sector and healthcare. Newly generated volumes of structured and unstructured big data will only continue to rise and road safety experts need to get a better grasp of it all if they are to leverage it appropriately for the public good. If not, the UK risks falling further behind other countries that have embraced large database linkage for the public benefit. A start has to be made using the official datasets already routinely collected, and often recognised as some of the highest quality in the world. There has been no greater impetus to link these data than exists now, a time when there is governmental drive, established mechanisms and an international movement to support quality linkage with a view to understanding society's problems better.

For some time there has been a standstill in the progress of moves to halt UK road deaths. Now innovation has to take centre stage, and the leveraging of rich datasets that are routinely curated must be one of the first steps. This apparent low-hanging fruit is not without its pitfalls, however. As this report has explained, when conducting work on sensitive health data, particularly linkage, a delicate balancing act between risk and benefit is needed. There is a risk that personal data might be used without consent against the wishes of the public, or, worse still, that members of the public could be identified and come to harm of some kind. On the other hand, there is a wealth of opportunity for important research on unlocked datasets which could provide huge insights into how challenges in road safety might be addressed. Of course, robust measures must be taken to ensure appropriate information governance and the safeguarding of confidentiality, but previous anxieties connected with linkage need to be conquered, and the gauntlet of administrative steps and regulatory requirements which have to be overcome must be run with optimism, as never before has there been such support for this work.

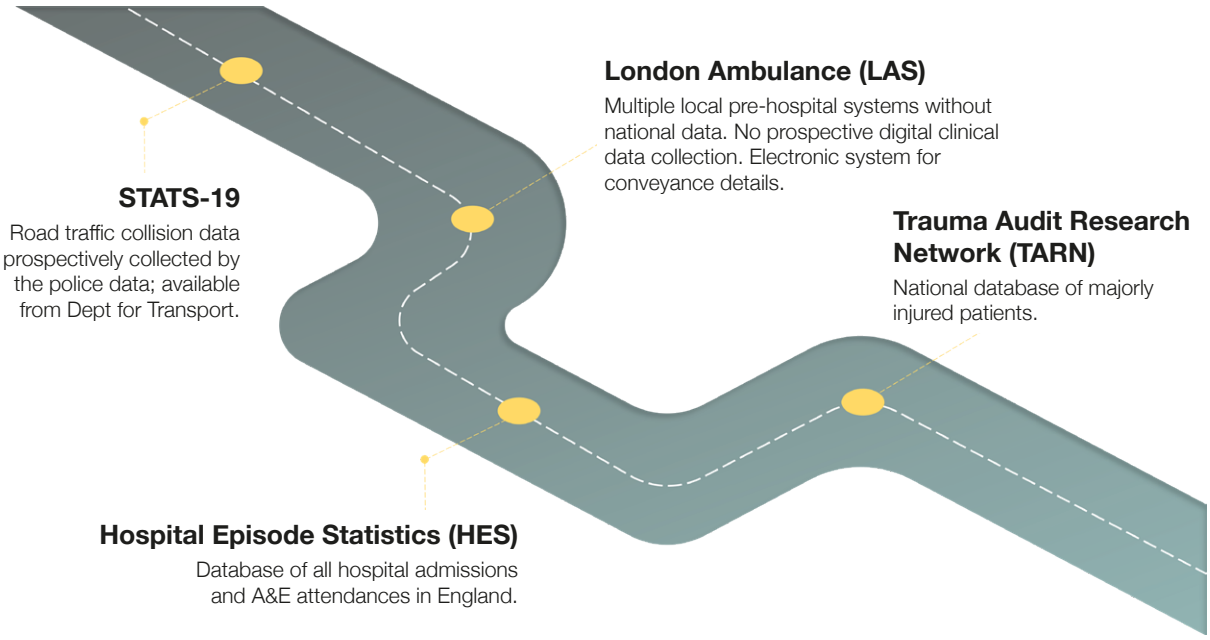
The RTI–AID project's exploratory findings on the linkage of public datasets relating to road safety in the UK – namely police reports in STATS19, ambulance records, and hospital data in the form of HES and TARN – suggest that such linkages are administratively, legally, ethically and politically feasible, provided there is multisectoral support in the form of a broad steering committee of transport and health professionals, and active engagement by data controllers and analysts from the parent organisations. Road safety advocates, researchers, policymakers and practitioners should be encouraged to come together and build on this further. Particularly if greater involvement of the public and patient groups can be enlisted, a national framework for linkage could be drawn up, creating a long-term basis for the data outputs. With this in place, a safely de-identified dataset describing the population experience from collision to discharge from hospital could be made available to researchers and policymakers to formulate interventions and monitor their efficacy.

In an age where information generation is growing exponentially, but promising technology and techniques are also emerging to tame it, the road safety community cannot lose its footing and miss the opportunity to learn what has not been possible to glean so far. There is no longer any place for compartmentalised thinking, and as a group road safety stakeholders must view transport and health as a continuum in road safety, whether it is in relation to how we investigate new phenomena, design policy or raise awareness. A drive for a national integrated dataset which captures this continuum would lay the foundation for a future with a joined-up world view – and constitute a momentous step that would go a significant distance in bringing about safer roads for all.

Appendix A Overview of National Road Safety Datasets

Within the United Kingdom, data on road collisions and injuries are collected by both government departments and the NHS. Each has different characteristics pertaining to origins of the data collected, parameters documented, publication time and accessibility (see Figure A.1). Here we discuss which organisations and sectors collect road safety data, and how this varies geographically across the nation.

Figure A.1: National datasets capturing elements of road safety from roadside to hospital



Source: Author's own

A.1 STATS19

In Great Britain specifically, national road accident statistics are derived from STATS19 data collected by the police and collated by the Department for Transport (DfT). STATS19 data is collected by the police when an incident occurs on a public highway involving at least one vehicle (this may be non-motorised such as pedal cyclists), and there is at least one person injured. However, not all incidents recordable in STATS19 meet the criteria for reporting to police under the Road Traffic Act 1988. Reportable traffic collisions under the Road Traffic

Act 1988 involve a mechanically propelled vehicle on public roads where there is injury to someone or something other than the driver or damage to another vehicle or property. Similarly, not all road collisions reportable by law meet the criteria for STATS19 (only those in which an injury is incurred).

The full dataset is not publicly available owing to the confidential nature of some fields, such as injury codes and home location postcodes; however, the remainder are accessible with custom downloads available via the DfT road traffic statistics website. For each collision, there are three elements recorded: collision circumstances (date and time, location, conditions), vehicles involved (vehicle and driver details), and casualties (injury severity, vehicle and relation to casualty, demographics). Since 2016 there were also changes to the way injury severity was recorded, with the previous codes of 'killed, serious or slight' being abolished in favour of injury-based reporting which it was hoped would remove some of the uncertainty that police officers were faced with on recording injury severity.

In Northern Ireland, the Department of Infrastructure collects road collision statistics as part of the Northern Ireland Road Safety Strategy.

A.2 Hospital Episode Statistics (HES)

The HES database contains details of all admissions, A&E attendances and outpatient appointments at NHS hospitals in England. The data is collected by healthcare providers during the admission and submitted to NHS Digital thereafter. Records contain information on patient demographics, clinical information (diagnoses and procedures), dates and means of admission, and location data including place of treatment and residence. Owing to the nature of the data, strict disclosure protocols apply on applying for bespoke HES datasets for research in order to preserve patient confidentiality. Preliminary data is available on a monthly basis, with finalised releases made annually, usually toward the end of the subsequent year.

The strength of HES data is that it captures all admissions and therefore all those seeking emergency care for road injury. However, limitations apply in both the quality of coding, whereby the original mechanism of injury may not be recorded, and the type of clinical data collected, which does not always make clear the true severity of injury. Even in the extremes of injury, a single operation may be life-saving and ultimately not require a protracted length of stay in hospital, whilst in others repeated smaller operations are needed prior to discharge. For such an admission, HES data may obscure the true severity of injury.

Outside England, hospital attendances and admissions are captured by similar datasets within each of the devolved nations. These are the Scottish Hospital Activity Statistics, Patient Episode Database for Wales and the Hospital Activity Statistics for Northern Ireland.

A.3 Trauma Audit Research Network

The Trauma Audit Research Network (TARN) is a national clinical audit for trauma care with data submitted by all hospitals in England, Wales, Northern Ireland and, more recently, the Republic of Ireland. Founded in 1990, it is the largest trauma registry in Europe and is funded by its feeding hospitals. It seeks to provide summary data at both national and local levels as to both injury prevalence and performance in injury care against established clinical standards. Still, not all trauma patients meet the criteria for inclusion. Those included must have suffered a significant injury as predefined by TARN, and have been admitted for at least three nights, admitted to critical care, or have died. This therefore leaves significant numbers of road traffic injuries outside the scope of TARN data collection. Requests for bespoke datasets are available upon application, with many Major Trauma Centres and Networks using the data for clinical research. The quality of data is variable and dependent on the resources available at the submitting hospitals, but internal checks are carried out periodically, including cross-validation with HES data. Corroborated data is usually available three months after it is submitted by hospitals.

In Scotland, data for the clinical management for the injured is collected by the Scottish Trauma Audit Group.

A.4 Pre-hospital data – the London Ambulance Service

The critical period between a collision and arrival at hospital is frequently omitted in both research and policy work. Pre-hospital data collection can be difficult to conduct owing to the chaotic environment, limited availability of practitioners to collect data at the same time as delivering emergency care and transporting patients, and poor data-gathering systems. Currently there is no systematic method of collecting pre-hospital data, or collating it at the national level. It is not routinely integrated with other health databases such as HES or TARN, although in some cases hospitals are able to incorporate pre-hospital data passed onto them by ambulance teams which is provided with their TARN submissions. Nonetheless, this data represents a critical link between the roadside and a definitive healthcare setting, providing information on initial clinical status, clinical progression and physical movement of the casualty within the Major Trauma Network of hospitals.

This fragmentation of pre-hospital services and data collection further drove the focus of the project on the area of Greater London, which is largely served by the LAS NHS Trust. The Trust is the busiest emergency ambulance service in the UK, answering two million 999 calls a year and attending over 3,000 emergencies daily. Paramedics collect broad parameters including call times, arrival and transport times, injuries identified, clinical status and interventions, and details of the final destination hospital. These are recorded on paper forms, duplicates of which are passed on to receiving hospitals. Data collection for research has therefore been performed manually or by restricting parameters analysed to the subset inputted into the digital record system, including emergency call and travel times, destination and arrival locations, and the clinical indication for call (e.g. road traffic injury).

These brief database descriptions demonstrate how rich data gathered on traffic collisions and consequent injuries within the UK really are. At the same time these data sources are far from integrated, with few mechanisms in place for sharing or linking data across the range of organisations collecting and hosting them.

Appendix B Further Details of Relevant UK Data Legislation

B.1 The General Data Protection Regulation (GDPR)

GDPR defines 'personal data' as information that relates to an individual. That individual must be identified or identifiable either directly or indirectly from one or more identifiers, or from factors specific to the individual, within the data.[61] Additionally, certain types of personal data are considered sensitive in nature and are under a higher degree of protection when it comes to processing. These 'special categories of personal data' include information relating to an individual's race, ethnicity, political opinions, religious beliefs, trade union membership, genetic data, biometric data, sexual orientation, sex life and health data.

GDPR and legal basis for processing personal data

There must be a legal basis for the processing of personal data according to Article 6(1) of GDPR. Processing shall be lawful only if and to the extent that at least one of the following applies:

a. the data subject has given consent to the processing of his or her personal data for one or more specific purposes;
b. processing is necessary for the performance of a contract to which the data subject is party or in order to take steps at the request of the data subject prior to entering into a contract;
c. processing is necessary for compliance with a legal obligation to which the controller is subject;
d. processing is necessary in order to protect the vital interests of the data subject or of another natural person;
e. processing is necessary for the performance of a task carried out in the public interest or in the exercise of official authority vested in the controller;
f. processing is necessary for the purposes of the legitimate interests pursued by the controller or by a third party, except where such interests are overridden by the interests or fundamental rights and freedoms of the data subject which require protection of personal data, in particular where the data subject is a child._

GDPR and legal basis for processing special categories of personal data

Processing of healthcare data must also be justified under Article 9 of GDPR which sets out the following potential conditions for legal processing:

a. explicit consent;
b. employment, social security and social protection (if authorised by law);
c. vital interests;
d. not-for-profit bodies;
e. made public by the data subject;
f. legal claims or judicial acts;
g. reasons of substantial public interest (with a basis in law);
h. health or social care (with a basis in law);
i. public health (with a basis in law);
j. archiving, research and statistics (with a basis in law).

B.2 The Data Protection Act 2018 (DPA)

The DPA incorporates additional requirements and safeguards into UK law beyond GDPR.

With respect to special categories of data, if using conditions (b), (h), (i), or (j), as is the case here, the work must also meet Part 1 of Schedule 1 of the DPA 2018:

Public health

3 This condition is met if the processing:

a. is necessary for reasons of public interest in the area of public health, and
b. is carried out:
    i. by or under the responsibility of a health professional, or
    ii. by another person who in the circumstances owes a duty of confidentiality under an enactment or rule of law.

Research etc.

4 This condition is met if the processing:

a. is necessary for archiving purposes, scientific or historical research purposes or statistical purposes,
b. is carried out in accordance with Article 89(1) of the GDPR (as supplemented by section 19)*, and
c. is in the public interest.

*Article 89(1):

> [1]Processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes, shall be subject to appropriate safeguards, in accordance with this Regulation, for the rights and freedoms of the data subject. [2]Those safeguards shall ensure that technical and organisational measures are in place in particular in order to ensure respect for the

principle of data minimisation. [3]Those measures may include pseudonymisation provided that those purposes can be fulfilled in that manner. [4]Where those purposes can be fulfilled by further processing which does not permit or no longer permits the identification of data subjects, those purposes shall be fulfilled in that manner.[76]

This means that to safeguard individuals in the datasets, research must be conducted on minimised data (i.e. data which contains no more information than that which is directly relevant and necessary for analysis to answer the research question). Additional safeguarding should also include pseudonymisation to reduce the possibility of re-identification of individuals during processing.

B.3 National Health Service Act 2006 and the Health Service Regulations 2002

Health data with potential identifiers not only counts as special category data under GDPR but also falls under the definition of confidential patient information as defined by section 251 of the National Health Service Act 2006, which applies to both the living and deceased.

Under the common law duty of confidentiality, when information is given in circumstances where it is expected that a duty of confidentiality applies, which is the case with health records and submissions to health datasets such as HES and TARN, information cannot usually be disclosed without patient consent.

This requirement can be waived where confidential patient information is to be used to benefit research, provided there is section 251 support. Section 251 of the National Health Service Act 2006 and its Regulations – Regulation 5 permits processing for a range of medical purposes, broadly defined to include preventative medicine, medical diagnosis, medical research, the provision of care and treatment and the management of health and adult social care services.

Section 251 recognises that there are essential health service activities such as research that require the use of confidential patient information when neither anonymisation nor retrospective consent are possible. In such cases section 251 permits the temporary lifting of the common law duty of confidentiality for medical purposes.[64]

Section 251 support is obtained through application to the Confidentiality Advisory Group (CAG), which acts under the Health Research Authority (HRA). CAG reviews applications for research and non-research work and determines whether there is sufficient public interest to temporarily lift the duty of confidentiality. Other bodies use CAG advice as the basis for their approvals, including HRA (research applications), the Secretary for State (non-research applications) and NHS England (data dissemination).

References

1. World Health Organization. Global Status Report on Road Safety 2023 [Internet]. Geneva: World Health Organization; 2023 Dec 13 [cited 2023 Dec 18]. Available from: https://www.who.int/teams/social-determinants-of-health/safety-and-mobility/global-status-report-on-road-safety-2023

2. World Health Organization. Road traffic injuries [Internet]. Geneva: World Health Organization; 2023 Sep 15 [cited 2023 Sep 20]. Available from: https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries

3. Lynam D, Nilsson G, Morsink P, Sexton B, Twisk D, Goldenbeld C. Sunflower: A comparative study of the development of road safety in Sweden, the United Kingdom, and the Netherlands. Leidschendam, The Netherlands: SWOV Institute for Road Safety Research; Crowthorne, Berkshire: Transport Research Laboratory TRL; Linköping: Swedish National Road and Transport Research Institute VTI; 2002.

4. Luoma J, Sivak M. Why is road safety in the U.S. not on par with Sweden, the U.K., and the Netherlands? Lessons to be learned. Eur Transp Res Rev. 2014;6:295–302. https://doi.org/10.1007/s12544-014-0131-7.

5. Hughes BP, Anund A, Falkmer T. System theory and safety models in Swedish, UK, Dutch and Australian road safety strategies. Accid Anal Prev. 2015;74:271–278. https://doi.org/10.1016/j.aap.2014.07.017.

6. Parliamentary Advisory Council for Transport Safety. Safe System. 2018 [cited 2023 Sep 17]. Available from: https://www.pacts.org.uk/safe-system/

7. Debyser A. Road safety in the EU. European Parliamentary Research Service; 2019. Available from: https://www.europarl.europa.eu/RegData/etudes/BRIE/2019/635540/EPRS_BRI(2019)635540_EN.pdf

8. Parliamentary Advisory Council for Transport Safety. UK Road Safety progress since 2010 ranks as one of the worst in Europe. PACTS [Internet]. 2021 Jul 27; [cited 2023 Sep 21]. Available from: https://www.pacts.org.uk/updated-results-for-gb-road-deaths-in-2020/

9. Vision Zero Network. What is Vision Zero? [Internet]. 2023 [cited 2023 Sep 17]. Available from: https://visionzeronetwork.org/about/what-is-vision-zero/

10. Action Vision Zero. Vision Zero - A History. [Internet]. 2022 [cited 2023 Sep 17]. Available from: https://actionvisionzero.org/resources/vision-zero-a-brief-history/

11. Bouwen L, Weijermars W, Johannsen H, Martensen H. Road Safety Thematic Report – Serious injuries. European Road Safety Observatory [Internet]. 2021 Jan; [cited 2023 Sep 21]. Available from: https://road-safety.transport.ec.europa.eu/system/files/2022-01/Road%20Safety%20thematic%20report%20Serious%20injuries_final.pdf

12. Department for Transport. STATS19 review: 2018 review [Internet]. London: Department for Transport; 2020 Mar [cited 2023 Sep 21]. Available from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/995117/stats19-review.pdf

13. Gill M, Goldacre MJ, Yeates DGR. Changes in safety on England's roads: analysis of hospital statistics. BMJ. 2006;333:73. https://doi.org/10.1136/bmj.38883.593831.4F.

14. Department for Transport. Estimating clinically seriously injured (MAIS3+) road casualties in the UK [Internet]. UK: Department for Transport; 2016 [cited 2023 Sep 20]. Available from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/556648/rrcgb2015-03.pdf

15. Department for Transport. Linking Police and Hospital data on Road Accidents in England: 1999 to 2009 results. February 2012. Available from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/995112/hes-linkage.pdf

16. International Transport Forum. Road Safety Annual Report 2021: The Impact of Covid-19 - Sweden. OECD Publishing, Paris. [Internet]. 2023 [cited 2023 Sep 17]. Available from: https://www.itf-oecd.org/sites/default/files/sweden-road-safety.pdf

17. Ministry of Transport, New Zealand. Sorted Study: Findings of the Study of Road Trauma Evidence and Data [Internet]. 2022 [cited 2023 Sep 17]. Available from: https://www.transport.govt.nz/assets/Uploads/SORTED2022Web.pdf

18. Hosseinzadeh A, Karimpour A, Kluger R, Orthober R. Data linkage for crash outcome assessment: Linking police-reported crashes, emergency response data, and trauma registry records. J Safety Res. 2022 Jun;81:21-35. https://doi.org/10.1016/j.jsr.2022.01.003.

19. Department for Transport. National statistics. Reported road casualties Great Britain, provisional results: 2022. 2023 [cited 2023 Sep 17]. Available from: https://www.gov.uk/government/statistics/reported-road-casualties-great-britain-provisional-results-2022/reported-road-casualties-great-britain-provisional-results-2022#:~:text=In%20reported%20road%20collisions%20in,of%2011%25%20compared%20to%202019

20. Statista. Statista Digital Economy Compass 2019 [Internet]. 2019. [cited 2023 Sep 17]. Available from: https://www.statista.com/study/52194/digital-economy-compass/

21. Big Data Framework. A short history of big data [Internet]. 2023 [cited 2023 Sep 17]. Available from: https://www.bigdataframework.org/knowledge/a-short-history-of-big-data/#_edn4

22. Statista. Worldwide data created [Internet]. 2023 [cited 2023 Sep 17]. Available from: https://www.statista.com/statistics/871513/worldwide-data-created/

23. Oracle. What is big data [Internet]. 2023 [cited 2023 Sep 17]. Available from: https://www.oracle.com/uk/big-data/what-is-big-data/

24. Educational Research Techniques. Big data data mining [Internet]. 2016 [cited 2023 Sep 17]. Available from: https://educationalresearchtechniques.com/2016/02/26/big-data-data-mining/

25. Munné, R. Big Data in the Public Sector. In: Cavanillas, J., Curry, E., Wahlster, W. (eds) New Horizons for a Data-Driven Economy. Springer, Cham. 2016. Available from: https://doi.org/10.1007/978-3-319-21569-3_11

26. Office for National Statistics. Developing standard tools for data linkage [Internet]. 2021 [cited 2023 Sep 17]. Available from: https://www.ons.gov.uk/methodology/methodologicalpublications/generalmethodology/onsworkingpaperseries/developingstandardtoolsfordatalinkagefebruary2021#main-approaches-to-data-linkage

27. Fellegi IP and Sunter AB, A theory for record linkage. J Am Stat Assoc, 1969. 64(328): p. 1183-1210.

28. Office for National Statistics. Data Linkage: Joining Up Data [Internet]. 2018 [cited 2023 Sep 17]. Available from: https://osr.statisticsauthority.gov.uk/wp-content/uploads/2018/09/Data-Linkage-Joining-Up-Data.pdf

29. Harron K. Data linkage in medical research BMJ Medicine 2022;1:e000087. https://doi.org/10.1136/bmjmed-2021-000087

30. Doidge J, Christen P, Harron K. Quality assessment in data linkage [Internet]. GOV.UK; [cited 2023 Sep 21]. Available from: https://www.gov.uk/government/publications/joined-up-data-in-government-the-future-of-data-linking-methods/quality-assessment-in-data-linkage

31. Bohensky, M.A., Jolley, D., Sundararajan, V. et al. Data Linkage: A powerful research tool with potential problems. BMC Health Serv Res (2010) 10, 346. https://doi.org/10.1186/1472-6963-10-346

32. Harron, K.L., Doidge, J.C., Knight, H.E., Gilbert, R.E., Goldstein, H., Cromwell, D.A., van der Meulen, J.H. A guide to evaluating linkage quality for the analysis of linked data. Int J Epidemiol (2017) 46(5):1699–1710, https://doi.org/10.1093/ije/dyx177.

33. Nesta. Councils and the data revolution [Internet]. 2023 [cited 2023 Sep 17]. Available from: https://www.nesta.org.uk/blog/councils-and-the-data-revolution-7-ways-local-authorities-can-get-more-value-from-their-data/

34. Epimorphics. 11 years since UKGovLD [Internet]. 2023 [cited 2023 Sep 17]. Available from: https://www.epimorphics.com/11-years-since-ukgovld/

35. Office for National Statistics. Joined-up data in government [Internet]. 2023 [cited 2023 Sep 17]. Available from: https://www.gov.uk/government/publications/joined-up-data-in-government-the-future-of-data-linking-methods/joined-up-data-in-government-the-future-of-data-linkage-methods#foreword-from-the-national-statistician-professor-sir-ian-diamond

36. McGrath-Lone L, Libuy L, Harron K, Jay MA, Wijlaars L, Etoori D, Lilliman M, Gilbert R, Blackburn R. Data Resource Profile: The Education and Child Health Insights from Linked Data (ECHILD) Database, International Journal of Epidemiology Feb 2022 Volume 51, Issue 1, Pages 17–17f, https://doi.org/10.1093/ije/dyab149

37. Ministry of Justice. Data First [Internet]. GOV.UK; [cited 2023 Sep 21]. Available from: https://www.gov.uk/guidance/ministry-of-justice-data-first

38. Administrative Data Research UK. Connecting administrative vehicle data for research on sustainable transport [Internet]. 2023 [cited 2023 Sep 17]. Available from: https://www.adruk.org/our-work/browse-all-projects/connecting-administrative-vehicle-data-for-research-on-sustainable-transport/

39. NHS Digital. Linked datasets supporting health and care delivery and research [Internet]. 2023 [cited 2023 Sep 17]. Available from: https://digital.nhs.uk/services/data-access-request-service-dars/linked-datasets-supporting-health-and-care-delivery-and-research

40. Ritchie F, Green E, Webber D, Mytton J, Deave T, Montgomery A, Woolfrey L, ul-Baset K, Chowdhury S. Enabling Data Linkage to Maximise the Value of Public Health Research Data: Summary [Internet]. Wellcome Trust; 2016 [cited 2023 Sep 21]. Available from: https://cms.wellcome.org/sites/default/files/enabling-data-linkage-to-maximise-value-of-public-health-research-data-summary-phrdf-mar15.pdf

41. Khan, M.N., Harris, M.L., Huda, M.N. et al. A population-level data linkage study to explore the association between health facility level factors and unintended pregnancy in Bangladesh. Sci Rep (2022) 12, 15165. https://doi.org/10.1038/s41598-022-19559-w

42. UK Health Data. Clinical Practice Research Datalink (CPRD) [Internet]. 2023 [cited 2023 Sep 17]. Available from: https://ukhealthdata.org/members/clinical-practice-research-datalink-cprd/

43. The Health Foundation. The Networked Data Lab [Internet]. 2023 [cited 2023 Sep 17]. Available from: https://www.health.org.uk/funding-and-partnerships/our-partnerships/the-networked-data-lab

44. Information Commissioner's Office. What is personal data? [cited 2023 Sep 17]. Available from: https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/personal-information-what-is-it/what-is-personal-data/what-is-personal-data/

45. Information Commissioner's Office. Lawful basis for processing: Special Category Data [cited 2023 Sep 17]. Available from: https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/lawful-basis/a-guide-to-lawful-basis/lawful-basis-for-processing/special-category-data/

46. Laurie G, Ainsworth J, Cunningham J, Dobbs C, Jones KH, Kalra D, Lea NC, Sethi N. On moving targets and magic bullets: Can the UK lead the way with responsible data linkage for health research? Int J Med Inform. 2015 Nov;84(11):933-40. https://doi.org/10.1016/j.ijmedinf.2015.08.011

47. Health Research Authority & INVOLVE. Public involvement in research: impact on ethical research 2016 [cited 2023 Sep 17].

48. Godlee, F. What can we salvage from care.data?. BMJ. (2016). 354. i3907. https://doi.org/10.1136/bmj.i3907

49. Sterckx S, Rakic V, Cockbain J, Borry P. "You hoped we would sleep walk into accepting the collection of our data": controversies surrounding the UK care.data scheme and their wider relevance for biomedical research. Med Health Care Philos. 2016 Jun;19(2):177-90. https://doi.org/10.1007/s11019-015-9661-6

50. Temperton J. NHS care.data scheme closed after years of controversy. Wired UK. 2016 [Internet]. 2023 [cited 2023 Sep 17]. Available from: https://www.wired.co.uk/article/care-data-nhs-england-closed

51. The Caldicott Committee, Department of Health. Report on the review of patient identifiable information. December 1997 Available from: https://webarchive.nationalarchives.gov.uk/ukgwa/20130124064947/http://www.dh.gov.uk/prod_consum_dh/groups/dh_digitalassets/@dh/@en/documents/digitalasset/dh_4068404.pdf

52. Crouch H. BMA and RCGP issue joint letter to NHS Digital over GPDPR programme. Digital Health [Internet]. 2021 June 7 [cited 2023 Sept 17]; [about 1 p]. Available from: https://www.digitalhealth.net/2021/06/bma-and-rcgp-issue-joint-letter-to-nhs-digital-over-gpdpr-programme/

53. Bradford L, Aboy M, Liddell K. COVID-19 contact tracing apps: a stress test for privacy, the GDPR, and data protection regimes. J Law Biosci. 2020 May 28;7(1):lsaa034. https://doi.org/10.1093/jlb/lsaa034

54. Wrigley D. Why has Palantir been given an interim contract to work on an NHS patient data project? BMJ 2023; 381 :p1482 https://doi.org/10.1136/bmj.p1482

55. RoadPeace. Crash not Accident Briefing Sheet [Internet]. 2022 [cited 2023 Sep 17]. Available from: [https://www.roadpeace.org/wp-content/uploads/2022/02/RP_Crash_not_Accident_Briefing_Sheet.pdf]

56. Journalists Reporting Guidance. Accidents vs collisions 2021 [Internet]. 2023 [cited 2023 Sep 17]. Available from: https://www.rc-rg.com/guidelines

57. Haddon W. The changing approach to the epidemiology, prevention and amelioration of trauma: the transition to approaches etiologically rather than descriptively based. Am J Public Health Nations Health. 1968 Aug;58(8):1431-8. https://doi.org/10.2105/ajph.58.8.1431.

58. Barnett DJ, Balicer RD, Blodgett D, Fews AL, Parker CL, Links JM. The application of the Haddon matrix to public health readiness and response planning. Environ Health Perspect. 2005 May;113(5):561-6. https://doi.org/10.1289/ehp.7491

59. Martensen H., G. Duchamp, V. Feypell, V. I. Raffo, F. A. Burlacu, B. Turner, and M. Paala. 2021. Guidelines for Conducting Road Safety Data Reviews. Washington, DC: World Bank." License: Creative Commons Attribution CC BY 3.0 IGO Available from: https://openknowledge.worldbank.org/server/api/core/bitstreams/bb8f4b8b-44dc-5e7e-ba46-930e78e956a0/content](https://openknowledge.worldbank.org/server/api/core/bitstreams/bb8f4b8b-44dc-5e7e-ba46-930e78e956a0/content

60. European Transport Safety Council. PIN Flash 16. Tackling the three main killers on the roads. A priority for the forthcoming EU Road Safety Action Program. (2014). Brussels: European Transport Safety Council. Available from: https://etsc.eu/an-overview-of-road-death-data-collection-in-the-eu-pin-flash-35/

61. Department for Transport. Reported road casualties in Great Britain: Frequently asked questions (FAQs) 2021 [cited 2023 Sep 17]. Available from: https://www.gov.uk/government/publications/reported-road-casualties-great-britain-frequently-asked-questions/reported-road-casualties-in-great-britain-frequently-asked-questions-faqs

62. Department for Transport. Linking Police and Hospital Data on Road Accidents in England:1999-2009. 2012 [cited 2023 Sep 17]. Available from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/995112/hes-linkage.pdf

63. Department for Transport. Reported road casualties Great Britain: 2015 annual report. Estimating clinically seriously injured (MAIS3+) road casualties in the UK. 2015 [cited 2023 Sep 17]. Available from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/556648/rrcgb2015-03.pdf

64. Cole, E. The national major trauma system within the United Kingdom: inclusive regionalized networks of care. Emergency and Critical Care Medicine June 2022 2(2):p 76-79. https://doi.org/10.1097/EC9.0000000000000040

65. Department for Transport. Linking STATS19 and TARN: an initial feasibility study. 2022 [cited 2023 Sep 17]. Available from: https://www.gov.uk/government/statistics/linking-stats19-and-tarn-an-initial-feasibility-study/linking-stats19-and-tarn-an-initial-feasibility-study

66. GitHub. RSS data TARN public [Internet]. 2023 [cited 2023 Sep 17]. Available from: [https://github.com/department-for-transport-public/RSS_data_TARN_public]

67. Nunn J, Barnes J, Morris A, Petherick E, Mackenzie R & Staton M. Identifying MAIS 3+ injury severity collisions in UK police collision records, Traffic Injury Prevention, (2018) 19:sup2, S142-S144, https://doi.org/10.1080/15389588.2018.1532205

68. Cambridge & Peterborough Road Safety Partnership. Targeting road injury prevention (TRIP) project report [Internet]. 2021 [cited 2023 Sep 17]. Available from: https://www.cprsp.co.uk/news/targeting-road-injury-prevention-trip-project-report

69. docbike.org Reducing Avoidable Harm in Motorcyclists through Injury Prevention & Roadside Critical Care. 2016 [cited 2023 Sep 17]. Available from: https://docbike.org/wp-content/uploads/2017/11/DocBike-national-strategy-July-2017-final.pdf

70. Department for Transport. Road accident in-depth studies (RAIDS) 2013 [cited 2023 Sep 17]. Available from: https://www.gov.uk/government/publications/road-accident-investigation-road-accident-in-depth-studies/road-accident-in-depth-studies-raids

71. Baker CE, Martin P, Wilson MH, Ghajari M, Sharp DJ, The relationship between road traffic collision dynamics and traumatic brain injury pathology, Brain Communications, Volume 4, Issue 2, 2022, fcac033, https://doi.org/10.1093/braincomms/fcac033

72. Cisco Annual Internet Report (2018–2023) White Paper. 2020 [cited 2023 Sep 17]. Available from: https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html

73. World Health Organization. Drinking water [Internet]. Geneva: World Health Organization; 2023 Sep 13 [cited 2023 Sep 20]. Available from: https://www.who.int/news-room/fact-sheets/detail/drinking-water

74. Beeharry MW, Moqeem K. The London Major Trauma Network System: A Literature Review. Cureus. 2020 Dec 9;12(12):e12000. https://doi.org/10.7759/cureus.12000.

75. Information Commissioner's Office. The UK GDPR. [cited 2023 Sep 17]. Available from: https://ico.org.uk/for-organisations/data-protection-and-the-eu/data-protection-and-the-eu-in-detail/the-uk-gdpr/

76. gdpr-info.eu. Article 89 GDPR Safeguards and derogations relating to processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes. [cited 2023 Sep 17]. Available from: https://gdpr-info.eu/art-89-gdpr/

77. Health Research Authority. Confidential patient information and the regulations. 2023. [cited 2023 Sep 17]. Available from: https://www.hra.nhs.uk/about-us/committees-and-services/confidentiality-advisory-group/confidential-patient-information-and-regulations/

78. NHS Digital. Data Access Request Service (DARS): process. 2023 [cited 2023 Sep 17]. Available from: https://digital.nhs.uk/services/data-access-request-service-dars/process

79. Langley JD, Dow N, Stephenson S, et al Missing cyclists Injury Prevention 2003;9:376-379. https://doi.org/10.1136/ip.9.4.376

80. Gildea K, Hall D, Simms C. Configurations of underreported cyclist-motorised vehicle and single cyclist collisions: Analysis of a self-reported survey. Accid Anal Prev. 2021 Sep;159:106264. https://doi.org/10.1016/j.aap.2021.106264.

81. Office for Statistics Regulation. Data Sharing and Linkage for the Public Good. [Internet]. 2023 [cited 2023 Sep 17]. Available from: https://osr.statisticsauthority.gov.uk/publication/data-sharing-and-linkage-for-the-public-good/

82. International Transport Forum. Road Safety Annual Report 2021: The Impact of Covid-19 - United Kingdom. OECD Publishing, Paris. [Internet]. 2023 [cited 2023 Sep 17]. Available from: https://www.itf-oecd.org/sites/default/files/united-kingdom-road-safety.pdf

83. Aitken, M., de St. Jorre, J., Pagliari, C. et al. Public responses to the sharing and linkage of health data for research purposes: a systematic review and thematic synthesis of qualitative studies. BMC Med Ethics 17, 73 (2016). https://doi.org/10.1186/s12910-016-0153-x

84. Goldacre, B & Morley, J. (2022). Better, Broader, Safer: Using health data for research and analysis. A review commissioned by the Secretary of State for Health and Social Care. Department of Health and Social Care. [cited 2023 Sep 17]; Available from: https://assets.publishing.service.gov.uk/media/624ea3788fa8f54a864cc6ba/summary-goldacre-review-using-health-data-for-research-and-analysis.pdf

85. UK Health Data Research Alliance, & NHSX. (2021). Building Trusted Research Environments - Principles and Best Practices; Towards TRE ecosystems (1.0). Zenodo. https://doi.org/10.5281/zenodo.5767586

86. Department of Health & Social Care. Data saves lives: reshaping health and social care with data. 2022 [cited 2023 Sep 17]. Available from: https://www.gov.uk/government/publications/data-saves-lives-reshaping-health-and-social-care-with-data/data-saves-lives-reshaping-health-and-social-care-with-data

87. World Medical Association. WMA Declaration on Ethical Considerations Regarding Health Databases. [Internet]. 2016 [cited 2023 Sep 17]. Available from: https://www.wma.net/policies-post/wma-declaration-of-taipei-on-ethical-considerations-regarding-health-databases-and-biobanks/

88. The International Population Data Linkage Network [Internet]. 2023[cited 2023 Sep 17]. Available from: https://ipdln.org/

89. Mitchell RJ, Cameron CM, Bambach MR. Data linkage for injury surveillance and research in Australia: perils, pitfalls and potential. Aust N Z J Public Health. 2014 Jun;38(3):275-80. https://doi.org/10.1111/1753-6405.12234.

90. Administrative Data Research UK. About ADR UK. [Internet] 2023 [cited 2023 Sep 17]. Available from: https://www.adruk.org/about-us/about-adr-uk/

91. Mason, L. (2018). A Comparison of Record Linkage Techniques. Bureau of Labor Statistics. https://www.bls.gov/osmr/research-papers/2018/pdf/st180060.pdf

92. Asher J, Resnick D, Brite J, Brackbill R, Cone J. An Introduction to Probabilistic Record Linkage with a Focus on Linkage Processing for WTC Registries. Int J Environ Res Public Health. 2020 Sep 22;17(18):6937. https://doi.org/10.3390/ijerph17186937

93. Farrow L, Evans J. Future registry research. Bone Joint Res. 2023;12(4):256-258. https://doi.org/10.1302/2046-3758.124.BJR-2023-0072

94. Health Data Research UK. Public and Participant Involvement in Population Research UK (PRUK): A scoping review of different approaches and recommendations.[Internet]. 2021 [cited 2023 Sep 17]. Available from: https://www.hdruk.ac.uk/wp-content/uploads/2021/12/Annex-2-Public-and-Participant-Involvement-in-PRUK-scoping-review.pdf

95. Aleia Clark Fobia, Jessica Holzberg, Casey Eggleston, Jennifer Hunter Childs, Jenny Marlar, Gerson Morales, Attitudes towards Data Linkage for Evidence-Based Policymaking, Public Opinion Quarterly, Volume 83, Issue S1, 2019, Pages 264–279, https://doi.org/10.1093/poq/nfz008

96. Cross L, Carson LE, Jewell A, Heslin M, Osborn D, Downs J, Stewart R. Guidance for researchers wanting to link NHS data using non-consent approaches: a thematic analysis of feedback from the Health Research Authority Confidentiality Advisory Group. Int J Popul Data Sci. 2020 Oct 2;5(1):1355. https://doi.org/10.23889/ijpds.v5i1.1355.

97. King's College London. Data linkages animation explores the evolution of healthcare records in research [Internet]. 2023 [cited 2023 Sep 17]. Available from: https://www.kcl.ac.uk/news/data-linkages-animation-explores-the-evolution-of-healthcare-records-in-research

98. Clark SJ, Halter M, Porter A, Smith HC, Brand M, Fothergill R, Lindridge SJ, McTigue M, Snooks H. Using deterministic record linkage to link ambulance and emergency department data: is it possible without patient identifiers? A case study from the UK. Int J Popul Data Sci. 2019;4(1):1104. https://doi.org/10.23889/ijpds.v4i1.1104.

99. NHS England. Ambulance Data Set. [Internet]. 2022 [cited 2023 Sep 17]. Available from: https://www.england.nhs.uk/urgent-emergency-care/improving-ambulance-services/ambulance-data-set/

RAC Foundation

Mobility · **Safety** · **Economy** · **Environment**

The Royal Automobile Club Foundation for Motoring Ltd is a transport policy and research organisation which explores the economic, mobility, safety and environmental issues relating to roads and their users. The Foundation publishes independent and authoritative research with which it promotes informed debate and advocates policy in the interest of the responsible motorist.

RAC Foundation
89–91 Pall Mall
London
SW1Y 5HS

Tel no: 020 7747 3445
www.racfoundation.org

Registered Charity No. 1002705

Designed and printed by
The Javelin Partnership Ltd
Tel: 0118 907 3494