Office for Statistics Regulation

Systemic Review Programme

Policing review phase two: Media Analysis

October 2019

Office for Statistics Regulation

We provide independent regulation of official statistics produced in the UK. Statistics are an essential public asset. We aim to enhance public confidence in the trustworthiness, quality and value of statistics produced by government.

We do this by setting the standards they must meet in the Code of Practice for Statistics. We ensure that producers of official statistics uphold these standards by conducting assessments against the Code. Those which meet the standards are given National Statistics status, indicating that they meet the highest standards of trustworthiness, quality and value. We also report publicly on system-wide issues and on the way, statistics are being used, celebrating when the standards are upheld and challenging publicly when they are not.

Acknowledgements

We would like to thank all the individuals and organisations that gave up their time to contribute to this Review as we gathered information and tested the ideas presented here. Their details are listed in Annex 1.

Contents

Foreword

At the Office for Statistics Regulation we have been looking at new ways to inform our work and support our vision that statistics serve the public good. This includes considering how we can make better use of a range of approaches to gaining insights from existing information which will help us understand how the public consume statistics and how statistics influence public debate.

This report focuses on the use of statistics around policing in the media and complements our first report on the use of statistics on policing in public debate, where we explored perceptions of public discourse about policing; how well statistics describe policing; and what might prevent statistics from better informing public understanding. It has allowed us to show how public conversations have changed and highlights potential gaps in the public debate and related statistics.

It is our first output in a series of trials to see how we can make better use of automation and technology to inform our work. It makes use of web-scraping and text analysis to gain insights into the media dialogue on policing. While at this stage there are limitations to the conclusions we can draw (see Annex 2), it does move us a significant step forward in understanding relevant issues and developing our capability. The work offers additional insight, which complements our more traditional approaches to gathering information and can help us challenge our assumptions.

It is not our expectation that analytical work like that set out in this report will ever replace the valuable information we get from qualitative work, but we see huge value in developing our use of technology to support the work we do. The work undertaken to inform this report provides a platform to build on as we look to make more use of text analysis in a range of ways, including to inform our future work programme and understand how statistics influence people’s daily lives and individual decision-making processes.

We welcome your interest in this work and would appreciate any feedback.

M. Gregory

Mary Gregory

Deputy Director for Regulation

Introduction

Statistics add value when they support society’s need for information.

1.1. In 2018 we launched a review of the value of statistics on policing to public debate. We published a first report on the Use of statistics in public discourse: the example of policing statistics earlier this year(Phase one of the review). This second report covers phase two of the review and builds on our ambition to support the use of statistics in public debate.

1.2. Our first report drew on input from a range of stakeholders and a review of official statistics publications. Our research suggested that statistics on policing did not fully inform public debate. There are gaps in information to support public understanding of the demands placed on the Police and their changing role.

1.3. Given the importance of the media in the way the public consume information about policing we wanted to complement phase one of the review with analysis of the media. This is intended to develop our understanding of how use of statistics in the media informs public debate and where there are gaps in the information available to support public dialogue. This report outlines our work to analyse media stories and builds our understanding of the topics that are driving public discourse about policing, including how statistics inform those stories.

1.4. To support the review, we gathered and analysed over 2,500 media stories about policing, published between January 2018 and February 2019. We used articles published by Policing Insight, in its media monitoring section. Policing Insight is a website dedicated to news and analysis on the police and policing throughout the UK, it is subscription based and is accessed via a login. It provided us with a data set comprised of local and national stories about policing, including the demands they face and crimes they deal with. We recognise that there are some limitations to the data, such as, the potential bias in the Policing Insight website sample or topics being skewed by major events like the Gatwick drone affair. However, we consider it to be a valuable source of information to complement what we already know and give us a broader perspective.

1.5. We have analysed the media coverage to understand the topics informing public debate around policing and what part statistics play in it. The majority of articles written about policing and crime are at least partially based on figures either from official statistics or other sources.

What we found

Summary findings

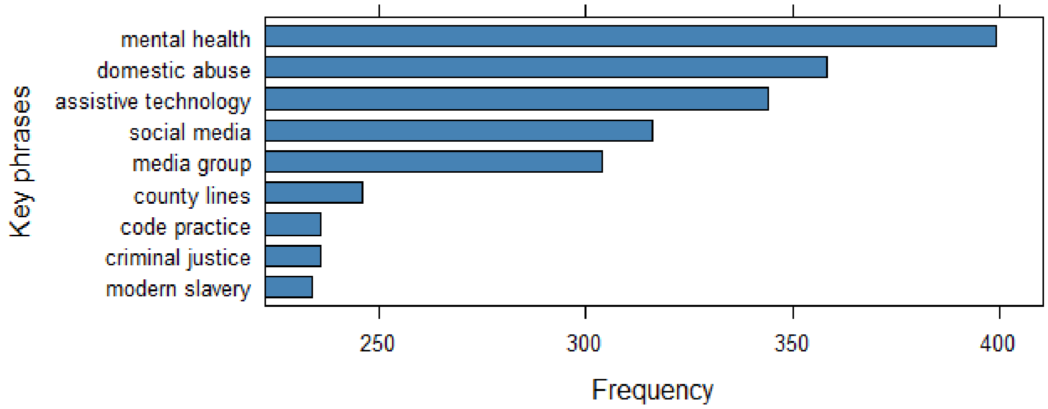
2.1. Through the topic analysis carried out as part of this review we were able to identify the most prevalent key words or phrases and topics gaining coverage in UK (national and local) media.

To support wider use of our analysis we have created an RShiny application (which works best in Google Chrome). The application is a trimmed down version of the dashboard we used to inform this report. The application allows users to interact with the data and produce different charts and analysis for their own use. For example, you can:

- Produce charts on the most common words associated with different news providers such as: The Daily Mail, Telegraph, The Liverpool Echo and The Herald;
- Investigate the most common three words associated with a word of your choice, for example statistics is associated with “official statistics publication”;
- Produce charts on words most commonly associated with each topic.

2.1. We looked at the frequency of noun and adjective phrases, in the media stories we gathered, see Figure 1. This allows us to understand the most common issues that are being raised in media related policing articles. From this summary it is not possible to see what is driving these phrases, but it does highlight how often certain issues are mentioned. Mental health is the most prevalent noun phrase. Our RShiny application allows anyone to investigate what is driving this. The phrase mental health is not traditionally associated with police work, however as our research showed in phase one, it is becoming more prevalent in policing vernacular. By looking at words associated with mental health in the relevant articles we can begin to get a sense of what is driving the stories. The RShiny application (sixth tab) shows common words associated with “mental” are incident, responding, services, broken and concerns. This suggests the police are focussing on this area and that mental health services are under strain - something we will consider in our systemic review of mental health statistics. Our understanding of the role the police play in protecting vulnerable people is limited; phase one research showed it is primarily anecdotal. The development of statistics in this area is important and the vulnerability section for the force management statements published by Her Majesty’s Inspectorate of the Constabulary Fire and Rescue Services HMICFRS show the kind of work being done to fill this gap.

Figure 1: Most frequently occurring noun phrases in the articles reviewed, January 2018 to February 2019.



2.2. Figure 1 also shows the broad range of most frequent topics highlighted, including domestic abuse, county lines and modern slavery. Following up the summary analysis with a deeper dive into the relevant articles provides the following insight into the key issues covered by each area:

- Mental health – refers to the way in which forces are starting to evolve prevention methods of dealing with vulnerable people, such as early intervention through councillors. There are also articles which discuss mental health for children while in custody, making sure the police have the correct procedures in place to help them, and articles highlighting the change police have faced in tackling mental health issues on the street, which has diverted them from fighting crime. All the articles illustrate how policing has evolved and is no longer just crime fighting, however the available statistics do not appear to have caught up with this change.

- Domestic abuse – many of the articles cover sexual abuse and harassment. There are also articles which look at honour based killings and some make reference to Karen Ingala Smith, the domestic abuse charity that gathers data on women killed as a result of domestic abuse. The latter is interesting as it highlights the use of non-official data to tell the story of domestic abuse.

- Assistive technology – refers to the development of technology to prevent public harm and using algorithms to make predictions about where to patrol.

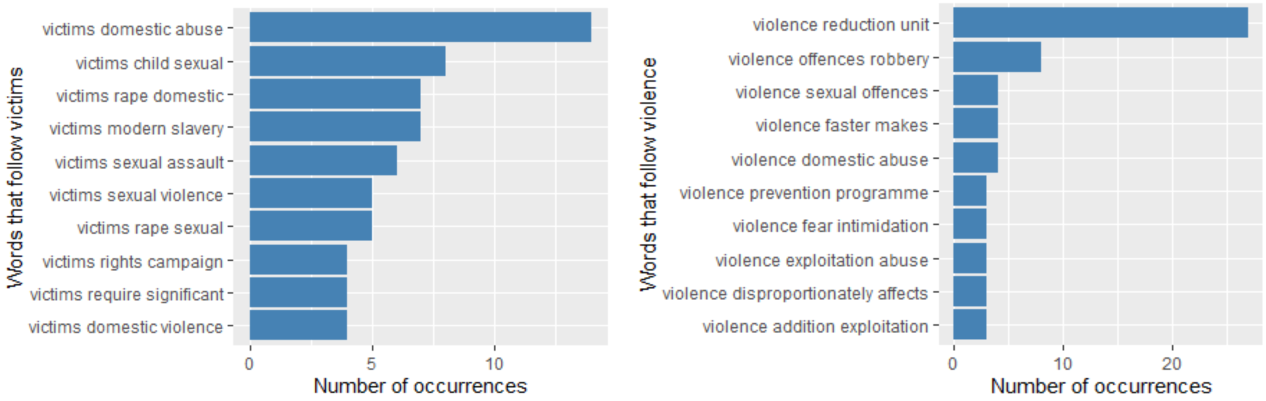
- Code practice – the articles that use this phrase reference police codes of practice and the code of practice for statistics. The split between these two areas is more focussed on police codes of practice which might be due to the specialised nature of the website we scraped. The code of practice for statistics is commonly mentioned when articles are discussing statistics, particularly police recorded crime.

- County lines – refers to the drug dealing networks which funnel drugs from cities to rural areas by exploiting children. Many of the articles concentrate on the nature of the issue rather than referring to any statistics or data. There are references to workforce strength and the capacity to deal with the county lines problem, illustrating that linking to police workforce statistics in relation to recorded crime would give users a more coherent picture of what the police are facing.

- Criminal justice – articles broadly discuss the financial strains the criminal justice system is under and how police are starting to use predictive analytics to determine were to deploy patrols in the face of these strains. There are also articles highlighting that there is a lack of confidence in the criminal justice system in relation to sexual assaults.

2.3. For greater insight we also identified words that appeared a lot in the articles we reviewed and looked at the words most likely to follow them. For example, the words victims and violence, see Figure 2. This helps us understand the common themes related to these words. For example, the highest combination of words following victims is domestic abuse, illustrating the importance of this subject within the public dialogue.

Figure 2: Most likely to follow the words victims and violence in the articles reviewed, January 2018 to February 2019.



a. victims

b) violence

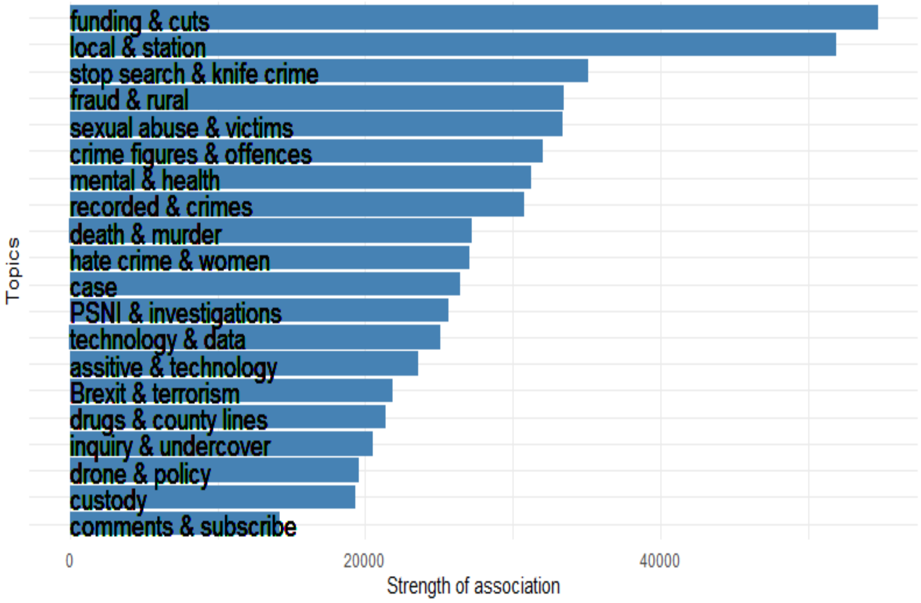
2.2. In addition to the analysis of key words and phrases set out above we undertook topic analysis (see annex 2). This analysis gives us a deeper understanding of the types of issues coming up in media stories by trying to find similar topics that come up in the different articles. We ran our model to find 20 topics (see annex 5). Each topic consists of words with similar meanings, see table 1. We set a limit to the number of similar words that were generated for each topic at five. We then used our prior knowledge from research during phase one to discern what the title of each topic would be.

Table 1: Top five words generated for each topic and corresponding topic name we gave, January 2018 to February 2019

. technology & data: technology, recognition, data, facial, legal
. drone & policy: drones, group, drone, registered, policy
. inquiry & undercover: surrey, comments, lawrence, undercover, inquiry
. stop search & knife crime: knife, violence, stop, search, violent
. Brexit & terrorism: brexit, deal, terrorism, salisbury, threat
. comments & subscribe: sure, comments, subscribe, join, days
. sexual abuse & victims: cases, sexual, abuse, evidence, victims
. recorded & crimes: crimes, recording, victims, recorded, reports
. funding & cuts: funding, services, budget, million, cuts
. hate crime & women: hate, comments, incidents, women, crimes
. PSNI & investigations: psni, federation, justice, investigation, belfast
. death & murder: death, family, attack, murder, back
. custody: house, custody, federation, latest, article
. assitive & technology: format, technology, assistive, help, gormley
. drugs & county lines: children, drugs, drug, county, lines
. crime figures & offences: figures, number, offences, crimes, drivers
. case: must, case, section, able, date
. local & station: local, contact, station, response, road
. fraud & rural: work, rural, fraud, communities, enforcement
. mental & health: health, mental, data, work, system

2.3. During phase one we found that much of the discussion on police demand and what the police do was influenced by a narrow set of statistics. Police recorded crime and the crime surveys were seen to be the primary source for supporting debate about policing itself. Our topic analysis correlated with our phase one research. Figure 4 shows the prevalence of each topic. Stop and search & knife crime, crime figures & offences and sexual abuse & victims appear to be generated from articles discussing crime survey statistics. This signifies that official statistics have great potential to inform and shape public debate and that statistics on policing as well as the police's output would allow the media to provide a more coherent picture.

Figure 4: The strength of association for each topic to all media stories we reviewed, ranked, January 2018 to February 2019



2.4. Figure 4 shows that the top five topics most closely associated to our scraped data are; funding & cuts, local & station, stop and search & knife crime, sexual abuse & victims and fraud & rural. Looking at these topics in relation to the articles provides greater insight and understanding about what they mean:

- Funding & cuts – many articles cover this topic in relation to the council tax precept rise for funding police forces, conveying disappointment in it not being large enough. The differences in funding are also highlighted between Scotland and England and Wales, with Scotland receiving its funding through the Barnett formula.

- Local & station – this topic is related to the local policing duties and the closing of stations in Scotland. Most articles convey that there has to be a local police presence in order to prevent the county lines issue getting more out of control.

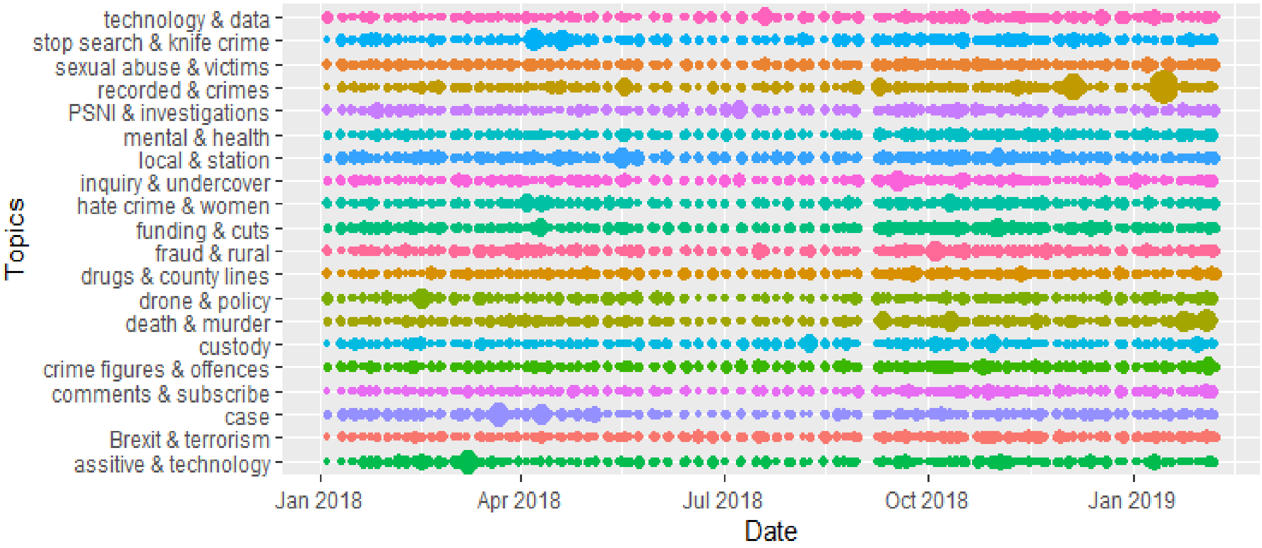
- Stop and search & knife crime – articles discuss knife crime in relation to police cuts and blame the rise in knife crime on cuts. The topic is also related to the victims and perpetrators of this type of crime being kids and the difficulty in criminalising children. Articles are broadly critical of stop and search procedures. There is a lot of coverage on this issue in Northern Ireland.

- Fraud & rural – several articles discuss internet fraud and sexual exploitation, which from phase one research we were aware was a growing concern and a gap in the current statistics. The rural element of this topic relates to fraud being committed due to the rise of broadband allowing gangs to connect with rural areas for the first time. Currently the most reliable source for understanding personal fraud in England and Wales is the Crime Survey of England and Wales. In Scotland there is ongoing discussion around recording crimes committed on the internet by perpetrators outside of the country. In both cases the current statistics are only partially telling the story of fraud.

- Sexual abuse & victims – articles discuss child sexual abuse enquiries some of which are historical cases. There are also references to modern slavery and the previously hidden nature of these crimes. There are some articles which reference Her Majesty's Inspectorate of Constabulary and Fire & Rescue Services (HMICFRS) investigations into the recording of these types of crime by police forces and how to improve the way in which they are recorded.

2.5. In order to understand the variation in prevalence of each topic over the last year we analysed the topics over time. Figure 5 is a jitter plot showing the prevalence of each topic over time. The larger or prominently clustered dots illustrate an increase in that topic's prevalence. Topics related to crime or recorded crime grow in prevalence during or near to when related official statistics are published. In contrast, for mental health and county lines there is a slow build up throughout the year, illustrating that discussion of these topics is not influenced by the publication of official statistics. The crime elements of the below topics are supported by official statistics, the other elements less so, illustrating a potential gap in our understanding of the demand faced by police.

Figure 5: Topic prevalence throughout the year, in all articles we reviewed, January 2018 to February 2019



- The topics; stop and search & knife crime, recorded & crime, crime figures & offences and hate crime & women grow in prevalence at similar periods throughout the year suggesting the debate around these topics is influenced by the Crime in England and Wales statistics. For example, this publication is released every four months; October, January, April and July. The level of prevalence in crime related topics increases slightly at similar points.
- The prevalence of the topics; mental & health and local & station remain constant through the period shown. While mental & health is self-explanatory, local & station less so. We investigated the words most closely associated with local and station. While there are associations to "local news" there are also associations to "funding", "sustainability" and "closure" of local police services. County & lines which was highlighted in our noun phrase analysis is also a topic which has been steady throughout the period shown. This has been an ongoing concern for police forces in England and Wales. The National Crime Agency (NCA) County Lines Drug Supply, Vulnerability and Harm report illustrates the work being done to bridge this gap and the collection of data on this subject.

Public Dialogue

3.1 The interviews and research we carried out in phase one of the review suggested that policing statistics and data do not fully inform the public across the UK about the current demands the police face, and how these demands are balanced in light of available resources. We have revisited some of the conclusions from the phase one report drawing on the new information we have gained from the media analysis.

3.2 In phase one we observed that recent public conversations about policing in England and Wales has often been linked to real, or perceived, increases in crime, especially violent crime. Our analysis did not provide clarity on this; we were able to look at the language the media use to describe crime to provide a better understanding of why this might be people's perception. Knife and recorded crime are discussed more regularly and the language around them is often embellished, with words such as; rising, wave, tide and alarming. The consequences for policing and the demands placed on them are lost, as the debate focuses on the nature of crime and the immediacy of the victims and perpetrators. As a result, the impact of the statistics to enhance debate is lost due to the type of language used, a more coherent statistical picture – one which doesn't predominantly rely on crime figures – would lead to a more informed public discourse, around what the police do and how their role is changing.

3.3 In Northern Ireland, during phase one we were told that one topic for recent ongoing debate has been about historic crimes committed during the Troubles and the balance of prosecutions related to deaths in each community. Other topics related to policing appearing in the media recently include resource implications of exiting the EU, paramilitary style attacks and use of police powers, specifically stop and search. Our analysis found that issues concerning the border and Brexit are prevalent topics. The most common topic to be generated from the Belfast Telegraph is around firearms and terrorism, with topics around funding and mental health also coming to light, while Derry Now and Belfast Live generate topics around Brexit, funding and youth crime. However, we were unable to pick up paramilitary attacks in our analysis.

3.4 In Scotland, during phase one we heard that media stories have recently included the issue of local versus centralised control including the closure of local police stations. This is part of an ongoing theme about Police Scotland that has also covered organisational issues such as the failure of a project to bring together separate IT systems into one and the management style of a previous Chief Constable. Topic analysis associated the local and centralised control debate with the recent move towards more digital working and the reduction of administrative tasks for police, which has resulted in less police stations. The ongoing themes about Police Scotland around organisational issues were also highlighted, with Philip Gormley the former Chief Constable, mentioned specifically, see annex 5.

3.5 In addition to these, we found recorded crime and domestic abuse are prominent topics, with recent discussions in Scotland around the crime rate being stable and the passing of the Domestic Abuse Act meaning that domestic abuse is now a crime. An investigation of the most common words associated with Scotland revealed rape and domestic abuse to be two of the most common, highlighting the importance of this debate in Scotland. The recent passing of Domestic Abuse Act will impact how the statistics will evolve as a result, and producers will have to ensure these changes are conveyed effectively in order to maintain effective public the debate.

3.6 The small number of journalists we spoke to as part of phase one were all interested in using statistics to evidence their stories and told us that the statistical evidence they would like to be able to use was not always readily available. As part of our analysis we looked at the frequency of different words which were commonly associated with Freedom of Information (FOI) requests and then searched for these terms in the data. We found the use of FOI requests to be quite frequent, but varied, ranging from FOI's asking for data on predictive policing algorithms to data on female genital mutilation (FGM) of young girls.

3.7 The range of questions being asked illustrates the importance of data and official statistics to the public debate, and the range of needs users have. However, we understand that FOI requests can often be burdensome on producers and that the information is sometimes available but difficult to find. The release of open data, including meta data, alongside statistical publications would help alleviate the burden of FOIs. Increasing the accessibility of the producer's websites so that they are easy to navigate could potentially lead to less questions being asked as well.

3.8 The analysis carried out in this section can be viewed in the Rshiny app we created to supplement the report.

Statistics and data on policing in the media

3.9 Although our analysis highlighted that the media is discussing issues around demands on policing, we found that the language used around topics related to the nature of crime gave these issues greater prominence. This means that a more nuanced debate about policing and the demands they face, is replaced with one which stresses the need for resource in certain areas over others.

3.10 Our analysis showed that the debate largely focusses on crime related topics; stop and search, knife crime, violent and recorded crime. The range of demands the police face outside of crime is missed, such as; mental health call outs, investigations and evidence building. Current media stories regarding funding and cuts to policing are framed within the context of rising violent crime, rather than the effects on police time and changes in society. There is an opportunity for producers of statistics to enhance the discourse by offering more coherent data, which more accurately reflects the range of police work.

3.11 One way of ensuring the debate remains engaged with what the police do, is to produce statistics and data that better inform the public and media. Our analysis produced topic areas with words like; knife, violent and stabbed, crimes, victims and recorded. Given the statistics published around these topics and based on our discussions in phase one, we can assume the public debate around crime demands on police is suitably informed. Despite police recorded crime no longer being National Statistics status, we acknowledge the work done by producers to ensure the public is effectively informed about crime related issues and the demand they place on the police.

3.12 There were areas highlighted in our analysis that the public are not informed on, such as; mental and health, hate crime & women. These are consistent with opinions from discussions in phase one that there are areas being discussed in media discourse, relating to police work, without data or statistics to underpin them. Further investigation of these words in context showed there are articles in our data that discuss the police being a service of last resort, diverting time from crime fighting to finding patients with mental health problems. These articles have little or no data or statistics with which to quantify the impact on policing and the police.

3.13 There is a reliance on Recorded Crime and the Crime Surveys of England & Wales (CSEW) to inform other parts of the debate. For example when looking at the words cuts and funding in context, they appear alongside articles which discuss the rise in violent crime rather than in relation to statistics on Police Funding for England and Wales. The topic funding & cuts does not increase in prevalence when the police funding publication is released, suggesting that the value of this publication is not being realised or there are limitations in the information being provided.

3.14 We proposed in our first report that there is a need to draw on other statistics and data, to inform the public debate. Our analysis suggests that the media is doing this. For example, one of the topics in England has been generated from inspections of police forces by HMICFRS. Analysis of commonly associated words shows British Transport Police to be a frequent combination of words and there are a number of articles using figures released by them. Specific surveys and analysis carried out by Mayor's Office for Policing and Crime (MOPAC) have also been used to evidence articles around stop and search procedures and hate crime. The use of these data are important to informing the public debate and illustrates that media outlets are trying to make use of facts to support their reports.

3.15 During phase one there was a broad consensus of the need to understand the time the police spent dealing with and doing different tasks. Our analysis highlighted some of the different work the police do in order to solve crime, such as evidence gathering and investigating. Current statistics only show the outcome of this type of police work, for example we see the recorded crime but have no understanding of the work that went into solving that crime. The provision of data to understand the time police spend on demands will be key in the public’s understanding of how policing is evolving from crime fighting into an organisation that deals with multiple social concerns.

3.16 Police Scotland are completing more work in order to understand the time police spend on different pieces of work and Scottish Government analysts are involved in early discussions on future developments. These statistics have the potential to fill a gap in the public’s understanding of what the police undertake as part of their role, and we look forward to seeing results of this work. Though work in these areas is just beginning, there are current opportunities for producers to improve statistics, such as; highlight statistics produced by non-official producers where relevant. ONS already provide links to the Crown Prosecution Service statistics on domestic abuse, which ensures greater coherence for users’ understanding of particular questions. This is a practice that should be employed more widely for statistics on policing and the nature of policing.

Conclusion and next steps

The evolution of policing and coherent statistics

4.1. We set out to identify the key areas of public debate around policing and to understand if the statistics supported the debate appropriately. The first phase of our research illustrated there is a strong consensus, for the need to improve police statistics. By analysing local and national media stories we have been able to supplement our understanding of the public debate around policing and more clearly understand the gaps in the statistics already produced.

4.2. While we understand the challenges presented to both the police and producers, phases one and two of our work taken together highlight the need for more coherent statistics. For example, to understanding how police use their time or how police finances link to outcomes. Producers of statistics play an important role in ensuring the public debate is well informed about policing and the demands the police face and we continue to encourage them to identify and fill gaps in existing information.

Next steps

4.3. As we continue to advocate the use of statistics in public dialogue, we want to continue the discussions we started in the first part of the review and invite others to join the conversation. As part of continuing this conversation we will:

- publish a follow up blog on Policing Insight, to generate further discussion at the local and national level;
- incorporate the analytical methods used for this report into our wider regulatory work programme to ensure our understanding of the public debate continues to grow;
- hold a round table between producers of statistics in the four nations in order to develop an understanding on presenting coherent policing statistics which support the public debate.

4.4. We will also work with organisations who do not produce official statistics, but who produce statistics and data that can inform public understanding about policing, to encourage voluntary adoption of the Code of Practice for Statistics with the framework of trustworthiness, quality and value at its core.

4.5. The use of statistics and data in the public discourse on policing will continue to grow. These steps will be part of a wider strategy aiming to ensure statistics effectively inform public debate.

Annex 1: Organisations that contributed to this review

Over the course of the review (phases one and two) we spoke to or received feedback from people in the organisations listed below.

Association of Police and Crime Commissioners

Audit Scotland

BBC

Dundee University, School of Social Sciences

Evening Standard

Her Majesty's Inspectorate of Constabulary, Fire and Rescue Services

Her Majesty's Inspectorate of Constabulary in Scotland

Home Office

House of Commons Library

London Mayor's Office for Policing and Crime

National Audit Office

National Crime Agency

National Police Chief's Council

Northern Ireland Policing Board

Northern Ireland Statistics and Research Agency

Office for National Statistics

Police Foundation

Police Ombudsman for Northern Ireland

Police Scotland

Police Service of Northern Ireland

Portsmouth University, Institute of Criminal Justice Studies

Queens University Belfast, School of Social Sciences

Scottish Government

Scottish Parliament Information Centre

Scottish Police Authority

Southampton University, Economic, Social and Political Sciences

Welsh Government

Annex 2: Methods for analysis

A2.1. We used articles published by Policing Insight, in its media monitoring section. Policing Insight is a dedicated website to news and analysis on the police and policing throughout the UK. We chose this website as it held a large collection of news stories going back to 2015.

Web scraping

A2.2. We used web scraping to gather the data for analysis. This technique allows data to be converted from an unstructured format – HTML tags – to a structured format. We used R to carry out the web scraping, specifically the packages Rvest and RSelenium. The latter allows you to scrape websites with JavaScript more easily.

A2.3. Before beginning our analysis, we prepared the data, primarily using Tidy Text. This package makes it easier to handle text data, providing a table that has one token (a word we wish to use for analysis) per row.

A2.4. We broke our text up into tokens and filtered out all the stop words – these are the most common words in a language i.e. or, and, if. We also filtered out the html and any words that were not useful in the analysis (this was a body of words we built up ourselves, for example words such as weather, sunny or cloudy).

Topic Analysis

A2.5. One of the issues with dictionary based or word frequency count approaches to text analysis, is that this method assumes a word has one meaning only. However, many words take on different meaning depending on the words around them.

A2.6. The technique we used for our analysis is called topic modelling. This method allows you to analyse groups of words together, instead of counting them individually. We are then able to understand the meaning of each word dependent on the broader context in which they were used.

A2.7. Topic models are mixture models, meaning that each observation is assigned a probability of belonging to a latent theme or topic. They also use iterative Bayesian techniques to determine the probability that each observation is associated with a given theme or topic. This means that the observations are initially given a random probability, but this becomes more accurate the more data that are processed.

A2.8. It is important to note that topic models are not a substitute for human interpretation, they allow us to make educated observations about how words contribute to different themes. For our analysis we used Latent Dirichlet Allocation which is one of the most common forms of topic modelling.

A2.9. LDA requires you to specify the number of topics the model should identify in the text. This is generally an arbitrary number and requires you to run the topic model several times, to see what is produced for each value. The STM package in R has a function which allows you to search for the best range of values to run your topic model on, providing multiple goodness of fit measures which help you to identify the best value to use. However, it should be noted that these measures are not perfect and human corroboration is the most appropriate way forward. We used a combination of both methods in order to ensure the best result.

A2.10. The model then randomly assigns each word to one of the *k* topics. The topic assignments are then iteratively updated based on the prevalence of the word across the topics. We stipulated the number of word assignments to be made to each topic as five.

A2.11. Topic modelling works best when the user has prior knowledge of the subject being modelled. Therefore, instead of writing code which automatically generated titles for each topic we used our prior knowledge to interpret the findings and categorise the topics manually.

Ngram Analysis

A2.12. An Ngram refers to a token, in this case a word. Tidy Text allows you to calculate the relationship between words and visualise this. It does this by splitting a piece of text into its separate words, which allows us to look at how often one word is followed by the next.

Grammar and key phrases

A2.13. The UDpipe package in R allowed us to tokenise our data and tag all parts of speech. This is based on a model that the package provides specifically for English. Using this package we were able to look for key phrases based on nouns and adjectives throughout the text.

A2.14. All the relevant code to conduct the analysis in this report and produce the RShiny app is located in annex 4.

Annex 3: List of R packages used

tidytext

dplyr

stringr

tidyr

tm

ggplot2

tidyverse

scales

ggthemes

drlib

plotly

UDpipe

Flexdashboard

Shiny

RSelenium

Rvest

Quanteda

stm

shinythemes

DT

statnet.common

reticulate

rsconnect

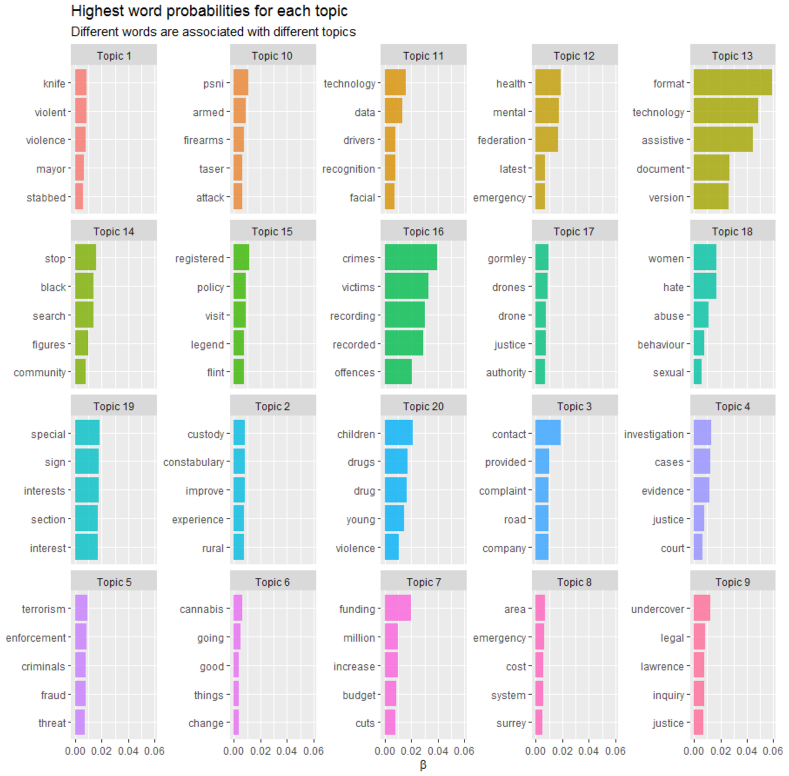
lattice

Annex 4: R code used and Shiny app

https://github.com/office-for-statistics-regulation/Policing-Review

Shiny app

Annex 5: Words that have the highest probability of making the topics



Highest word probabilities for each topic

Different words are associated with different topics