Part 1: Context – the benefits and challenges of using administrative data

Introduction

1.1 This report considers the risks surrounding the use of administrative data for statistical purposes. It identifies some examples of best practice across government in addressing those risks and presents some mechanisms for statisticians to implement when considering the quality of the data and the effect of any weaknesses on the derived statistical outputs. This report reviews: the quality checks that are carried out on administrative data before they are sent to a statistical producer; how they are questioned and examined; and how the inherent uncertainty in the data is communicated to the users of the statistics that are produced from them. The Authority recognises the resource challenges faced by statistical producers and advocates a proportional and pragmatic approach to the way that producers assess the level of assurance that is required.

1.2 This section presents the benefits and challenges of using administrative data in the production of official statistics. It then considers the weaknesses of administrative data and the role of quality assurance in addressing such limitations.

Use of administrative data in the compilation of official statistics

1.3 Administrative data are data collected for non-statistical purposes, for example, for registering births and deaths or administering benefits. It can often be personal information, for example, a person's hospital records. Administrative data can be considered as:

(i) registration records collected for an administrative purpose, and then compiled (in principle, automatically) to form a database of administrative data (for example, birth and death records)

(ii) those collected for operational purposes, such as, clinical records and payments of benefits. These can be subject to differing local administrative practices and therefore might be of variable quality, especially if those tasked with collecting the data do not have a full understanding of the end purpose for the data (for example, police recorded crime statistics).

1.4 The use of administrative data for the compilation of official statistics has many benefits, it can: achieve cost efficiencies in terms of re-using data; allow scheduled and timely collation of data from a large number of suppliers; and reduce the response burden.

Challenges in using administrative data for statistical purposes

1.5 However, there can be limitations in the nature of administrative or operational systems that can affect the statistics derived from the underlying data. Such problems may arise from differences in definitions preferred in the statistical and operational situations, as well as changes in the operational definitions and circumstances over time. A lack of standardisation in data collection procedures, IT systems and differing local policies and priorities, can also affect the statistics. These situations require investigation by statistical producers and clear communication about the limitations to users. Both data suppliers and statistical producers need to take account of public perceptions about the use of personal data for statistical purposes[2] and ensure that the data are sufficiently anonymised and secure. The computational (sorting, aggregating and linking data) and inferential (identifying whether change is real, or due to chance, or to poor data quality) challenges are striking and illustrate that these contemporary concerns are evolving and dynamic. In addition, in recent years there has been considerable interest in 'big data'[3] which reflects these issues on a vastly larger scale. Box A presents a series of challenges that producers commonly face when using administrative data in the production of official statistics.

Addressing uncertainty in the data

1.6 These challenges can affect different aspects of the quality of the data, such as the reported uncertainty around the data, as well as their comparability, standardisation and coherence and enabling the linkage with other datasets. Producers and users commonly recognise issues of uncertainty and bias in relation to survey-based statistics, and describe their scale by reporting measures such as sample size, response rates, measures of variance and precision, or descriptions of the likely sources of bias in relation to survey design and sampling. Quality measures collated during each stage of the survey process are used as the basis of an explanation for users about the quality of the statistics based on the survey data. In addition, bias may be assessed through comparison or linkage with other data sources. Less common, however, is the consideration of the inherent weaknesses in administrative or operational systems and their affect on statistics derived from them.

[2] Research carried out by ONS has revealed that the public expressed mixed opinions about the use of their public data for research and statistical purposes (http://www.ons.gov.uk/ons/about-ons/who-ons-are/programmes-and-projects/beyond-2011/beyond-2011-report-on-autumn-2013-consultation--and-recommendations/public-attitudes-report.pdf). Further research carried out for the Administrative Data Research Network identified that the public are concerned when administrative data are used by other agencies http://www.esrc.ac.uk/_images/Dialogue_on_Data_report_tcm8-30270.pdf

[3] 'Big data' typically refers to massive data sets which have the potential to reveal interesting or valuable insights into underlying processes and mechanisms which would not normally be apparent with smaller data sets. 'Big' can refer to the number of cases, the number of variables, the number of characteristics, the rate of data collection, or simply the complexity of the data.

Box A

Challenges using administrative data for statistical purposes:

Lack of standardised application of data collection:

- inconsistencies in how different suppliers interpret local guidance
- differences in the use of local systems for the intended administrative function
- the distortive effects of targets and performance management regimes
- differing local priorities, data suppliers might require higher levels of accuracy for certain variables (for example payments) but less so for other aspects that are important to the statistical producer (for example demographics)

Variability in data suppliers' procedures:

- statistical producers typically do not have direct control over the development of guidance for data entry
- local checking of the data can be variable and might not identify incorrect coding or missing values
- local changes in policy could impact on how the data are recorded or on the coverage of the statistics

Quantity of data suppliers:

- there can be a large number of data suppliers, often spread geographically
- there can be many data collectors providing their data to an intermediary organisation for supply to a statistical producer

Complexity and suitability of administrative systems:

- administrative datasets can be complex containing large numbers of variables; it takes time, and therefore resource, to extract the necessary data required by the statistical producer
- data collation can be hampered by IT changes at the data supplier level
- data might need to be manipulated by the data supplier to meet the structural requirements of the statistical producer, leading to potential for errors

Public perceptions:

- lack of knowledge about use of personal data for statistical purposes
- concern that personal data should be sufficiently anonymised and secured

Quality assurance

1.7 Quality management encompasses the full range of activities carried out by statistical producers in the production of official statistics, from the initial design of data collection through to the dissemination of the statistics. A critical element of this is 'quality assurance', defined as 'the part of quality management focused on providing confidence that quality requirements will be fulfilled'[4]. Traditional quality assurance activities, such as reviewing trends or comparing data across regions, can provide statistical producers with indications of where further

[4] *International Organization for Standardization (2005):* Quality management systems – Fundamentals and vocabulary (ISO 9000:2005). http://www.iso.org/iso/catalogue_detail?csnumber=42180

investigation of the underlying data could be required. Post-collection quality assurance methods, such as data validation, are an essential part of the quality assurance process, but can be of limited value if the underlying data are of poor quality. The quality of the entire statistical process directly affects the statistical products. While statisticians have demonstrated some appreciation of the limitations of administrative data, and in some cases developed good quality assurance processes after they receive the data, there has been a lack of application of critical judgment of the underlying data from administrative systems *before* the data are extracted for supply into the statistical production process. As with survey data, producers need to: investigate the administrative data to identify errors, uncertainty and bias in the data; make efforts to understand why these errors occur and to manage or, if possible eliminate, them; and communicate to users how these could affect the statistics and their use. The Authority recognises that there are certain circumstances in which regular, systematic audit of the underlying data is essential to increase both the quality of, and public confidence in, statistics produced from administrative data.

Audit

1.8 Audit should be a key part of the administrative data quality assurance process. In this context audit means an examination of records to check their accuracy and it includes inspections and other reviews by 'neutral internal or external experts'[5]. Administrative data underpinning official statistics can be subject to, or feature in, various kinds of audit, depending on their operational context, for example: financial, clinical, social care and statistical audit in which a sample of existing cases is investigated. These activities might be conducted on behalf of the data supplier bodies themselves as internal audit, or for regulators, such as the Care Quality Commission (CQC), or external audit or formal inspection regimes for example by the National Audit Office (NAO) or HM Inspectorate of Constabulary (HMIC). These audits should supplement, but not replace, detailed quality assurance checks carried out by statistical producers. The findings from reviews of audit arrangements will not necessarily lead to quantitative estimates of quality but can provide a richer body of evidence to inform judgments about:

- the suitability of the administrative data for use in producing official statistics
- factors the statistical producer needs to take into account in producing the official statistics
- the information that users need to know in order to make informed use of the statistics.

[5] ESS Data Quality Management Tools paper:
http://epp.eurostat.ec.europa.eu/portal/page/portal/quality/quality_reporting

Existing guidance

1.9 Guidance already exists for producers about the use of administrative data for statistical purposes[6] and the National Statistician's Office has recently circulated interim guidance[7] for producers about how to consider more carefully the quality of administrative data. In addition, there is a range of documentation available from Eurostat and some development of this topic by National Statistics Institutes (see Annex A). This review builds on this existing work.

The Authority's evaluation guide

1.10 This report highlights (in Part 2) some practices that we have identified from across the Government Statistical Service and some lessons learnt that can aid other statistical producers (fuller information is provided in Annex C). We then present evaluation guidance for statistical producers, to aid them in developing a better understanding about the quality of administrative data (in Part 3). The Authority recognises that producers are operating under tight resources; a critical aspect of addressing the concerns outlined in the paper is that statisticians take a proportionate approach based on the degree of concern about the quality of the underlying data and the public interest in the statistics – that is, the types of decisions that are informed by the statistics. The mechanisms presented in Part 4 provide statisticians with guidance on how to make these appropriate judgments. Part 5 specifies the relevant practices in the *Code* and the Authority's expectations for compliance. We conclude that section by highlighting three recommendations for statistical producers using administrative data to produce official statistics.

[6] NSO Guidance, Use of Administrative or Management Information: https://gss.civilservice.gov.uk/blog/2014/05/interim-administrative-data-guidance/
[7] https://gss.civilservice.gov.uk/wp-content/uploads/2012/12/Interim-Admin-Data-guidance.pdf