Public Health & Intelligence

Assuring the Quality of Data and Enhancing Trustworthiness and Public Value of Statistical Outputs

Document Control	
Version	Version 0.3
Date Issued	1 May 2017
Authors	Richmond Davies & Jill Ireland
Comments to	NSS.nssstatsgov@nhs.net

Version	Date	Comment	Author
0.1	07/11/2016	1st draft of paper, following preliminary discussion at Statistics Advisory Group (SAG) Meeting.	Richmond Davies & Jill Ireland
0.2	10/04/2017	2nd draft for presenting at SAG meeting of 13/04/2017	Richmond Davies & Jill Ireland
0.3	01/05/2017	3rd draft incorporating amendments following scrutiny at SAG meeting	Richmond Davies & Jill Ireland

BACKGROUND

The Code of Practice for Official Statistics[1] sets out the necessary principles and practices to produce statistics that are trustworthy, of high quality and of public value. Compliance with the Code is a statutory requirement on NHS National Services Scotland.

The United Kingdom Statistics Authority[2], in January 2015, published a toolkit[3] for administrative data quality assurance. The toolkit explains the nature of the regulatory standard that the Office for Statistics Regulation (the regulatory arm of the Authority) will apply when determining the suitability of statistical practices used by statistical producers for the quality assurance of administrative data. Fully complying with the toolkit helps organisations meet the highest standards of quality and trustworthiness.

[1] https://www.statisticsauthority.gov.uk/monitoring-and-assessment/code-of-practice/
[2] https://www.statisticsauthority.gov.uk/
[3] https://www.statisticsauthority.gov.uk/wp-content/uploads/2015/12/images-qualityassurancetoolki_tcm97-44368.pdf

Data quality is defined in the context of the users of the data. Data are considered to have sufficient quality or to be 'fit for purpose' depending on their intended purpose. The purpose may include management information for operational use, decision making, local reporting or statutory reporting which will require increasing levels of data quality. Where data quality issues are identified, it is expected that they should be reported as metadata clearly to the users including their impact and any plans for improving the quality.

> **The quality of a statistical product can be defined as the fitness for purpose of that product. More specifically, it is the fitness for purpose with regards to the following 5 quality dimensions: relevance, accuracy and reliability, timeliness and punctuality, accessibility and clarity, and coherence and comparability[4].**

This paper presents the data quality assurance and statistical output quality considerations for analysts in NHS National Services Scotland.

QUALITY CONSIDERATIONS FOR RAW DATA

Analysts should understand any issues which impact on the statistical output. In order to establish the foundations, an understanding is needed of the whole data life cycle from collection to capture and storage, to cleaning and analysis, to dissemination and feedback. The more you understand your data and its origins, the more likely you are to identify any issues that need to be investigated and improved upon.

The data management teams have a good understanding about the source, completeness, accuracy, frequency of submission and data issues at source as evidenced by their monitoring and quality assurance activities and interactions with analytical teams. However, to comply with good practice and recommendations by the UK Statistics Authority, there should be at least a basic understanding of collection, submission and quality matters by the data management and analytical staff who interact with the raw data so that their use of the data can be better informed by the context. Analytical teams are expected to be curious about the origin of the data they use in analysis and liaise more with data quality assurance and data monitoring teams who are very experienced in working with data suppliers and understanding their local issues.

Staff should have at least a basic understanding of the data source as follows:

- What data are collected
- How the data are collected
- Whether the collection method has an impact on data quality
- Those responsible for collecting the data at source
- The training provided to data entry personnel
- The information governance processes that are in place
- The checks that are undertaken to ensure and monitor the accuracy and completeness of the data at source

Analysts should have good links and communicate regularly with their data management colleagues in order to be reassured of the following:

- The process and frequency of data submission to NSS

[4] https://www.ons.gov.uk/methodology/methodologytopicsandstatisticalconcepts/qualityinofficialstatistics

- The communication channels with suppliers: i.e. the nature of the meetings that take place between data management and the data suppliers and whether or not analysts are involved
- The key data management contacts for the datasets of interest
- Whether or not data management colleagues or other analysts track data accuracy and completeness
- Data coverage issues and potential sources of bias in the data collection and supply process

On receipt of data, or on retrieving the data, staff should, if appropriate, engage in data profiling for analysis of individual attributes and data mining for discovering specific patterns in the dataset in order to understand whether the data is complete and whether there are other data issues. Table 1 below can be used as a guide:

Table 1: *Data profiling and data mining activities to discover problems*

Problem	Metadata	Examples
Illegal values / Invalid characters	Cardinality	If male =1, female=2, other=3, then cardinality >3 indicates a problem
	Maximum, minimum	Maximum and minimum should be not be outside the permissible range
	Variance, deviation	Variance, deviation of statistical values should not be higher than threshold
	Dates	If date of discharge is earlier than date of admission for a single episode of care Invalid date values
	String pattern	Phone numbers of GP Practices with erroneous/missing area codes
	Alphanumeric values	Conflicts in values: 'city' and 'postcode' do not correspond
Misspellings	Attribute values	Sorting on values often brings misspelled values next to correct values
Missing values	Null values	Percentage/number of null values
	Attribute values and default values	Presence of default value may indicate real value is missing
Duplicates	Attribute values	Sorting values by number of occurrences; more than one occurrence my indicate duplicates
Inconsistent aggregating	Attribute values	Column values do not sum, when they should

It is important to provide feedback on data quality issues to data management colleagues or direct to the data suppliers if data management colleagues are not involved in the data life cycle process.

DETERMINING THE QUALITY OF THE STATISTICAL OUTPUT

Staff should be aware of the public interest in the statistics they produce as well as the risk of data quality concerns. Together, they will guide the staff in determining the level of data quality assurance required for the statistical output.

According to the quality assurance toolkit, the critical judgment about the suitability of the administrative data for use in producing official statistics should be pragmatic and proportionate, made in the light of an evaluation of the likelihood of quality issues arising in the data that may affect the quality of the statistics, and of the nature of the public interest served by the statistics.

The staff should first assess the risk of data quality concerns as shown in table 2 below:

Table 2: *Risk profile of data quality concerns*

Risk profile of data quality concerns	Description
Low risk of data quality concerns	Clear agreement about what data will be provided by the supplier, when, how, and by whom; when there is a good appreciation of the context in which the data are collected, and the analyst accepts that the quality standards being applied meet the statistical needs
Medium risk of data quality concerns	High risk factors have been moderated through the use of safeguards such as operational checks, and effective data supplier/NHS NSS communication arrangements. Also, the use of other data types such as surveys and census data are used in the statistical output
High risk of data quality concerns	Many data supplier bodies are involved including intermediaries and where complex data collection processes have limited independent verification and oversight

The staff should then assess the public interest profile of the statistics as shown in table 3 below:

Table 3: *Public interest profile of the statistics*

Public interest profile of the statistics	Description
Low profile statistics	Politically neutral subject; interest limited to niche user base, and limited media interest.
Medium profile statistics	Wider user and media interest, with moderate economic and/or political sensitivity.
High profile statistics	Economically important, reflected in market sensitivity; high political sensitivity, reflected by parliamentary committee hearings; substantial media coverage of policies and statistics; important public health issues; collection required by legislation.

Finally, the staff should use the risk profile of data quality concerns and the public interest profile level to determine the level of assurance that the statistical output must meet. The levels of assurance are:

A1 = Basic assurance: statistical producer has reviewed and published a summary of the administrative data QA arrangements

A2 = Enhanced assurance: statistical producer has evaluated the administrative data QA arrangements and published a fuller description of the assurance

A3 = Comprehensive assurance: statistical producer has investigated the administrative data QA arrangements, identified the results of independent audit, and published detailed documentation about the assurance and audit

The risk/profile matrix shown in table 4 should be used to determine the level of assurance:

Table 4: *Risk/Profile Matrix*

Level of risk of quality concerns	Public interest profile		
	Low profile statistics	Medium profile statistics	High profile statistics
Low risk of data quality concern	Statistics of lower quality concern and lower public interest [**A1**]	Statistics of low quality concern and medium public interest [**A1** or **A2**]	Statistics of low quality concern and higher public interest [**A1** or **A2**]
Medium risk of data quality concern	Statistics of medium quality concern and lower public interest [**A1** or **A2**]	Statistics of medium quality concern and medium public interest [**A2**]	Statistics of medium quality concern and higher public interest [**A2** or **A3**]
High risk of data quality concern	Statistics of higher quality concern and lower public interest [**A1**, **A2** or **A3**]	Statistics of higher quality concern and medium public interest [**A3**]	Statistics of higher quality concern and higher public interest [**A3**]

Levels A1, A2 and A3 assurance exist across each of the following four areas of activity:

- Administrative data collection
- Communication with data suppliers
- Data suppliers' QA principles, standards and checks
- Statistical producers' QA investigations and documentation

Descriptions of each of the three levels of assurance across each of the 4 areas of activity are shown in administrative data quality assurance matrix in appendix 1. Please contact the Statistical Governance Team at nss.NSSstatsgov@nhs.net for any issues related to your interpretation of the three levels of assurance.

PUBLIC VALUE AND TRUSTWORTHINESS

During the analysis of the data and prior to disseminating any statistical output, staff should ensure that:

- The PHI Checking Guidance[5] has been adhered to
- Any syntax or algorithm used for performing checks should contain appropriate commentary and should be documented
- There are no major unexplained differences in figures between previously published years
- Correct version of SIMD is used. Contact NSS.isdGPD@nhs.net if any queries
- Correct Population Estimates are used. Contact NSS.isdGPD@nhs.net if any queries
- Correct matching of data to postcodes is done. Contact NSS.isdGPD@nhs.net if any queries
- The most recent versions of lookup files are used. Contact NSS.isdGPD@nhs.net if any queries
- Definitions are consistent over time, or any differences are explained
- The output, if appropriate, contains in-depth analysis and commentary, rather than just descriptive statistics and that a summary of the methodology is provided to aid the reader
- Analysts provide an explanation as to why they feel the statistics are robust for use

To add value for your users and increase the trustworthiness of the output:

- Summarise strengths and communicate any limitations relating to particular uses of the statistics
- Communicate uncertainty, where appropriate, using confidence intervals
- Convey clearly your judgement and the key evidence that helps inform that judgement
- Use plain English, where possible, but particularly so for the main points
- Help users visualise the data e.g. using simple diagrams and charts and, where appropriate, static or interactive dashboards
- Provide accompanying metadata to ensure that the output variables, caveats and issues are understood by the end users, and continue to be accessible into the future
- Identify your users, listen to their feedback and modify your outputs to meet their needs, if possible, while being pragmatic and open about what you can realistically modify based on your available resources

[5] http://www.isdscotland.org/About-ISD/Methodologies/_docs/PHI_CHECKING_GUIDANCE_20161012.pdf

- Based on patterns of information requests from customers, proactively publish as much as you can (e.g. frequent three star open data compliant CSV and accompanying metadata file updates which are supplemented with annual in-depth reports)

Principle 4, Practice 2 of the Code:
'Ensure that official statistics are produced to a level of quality that meets users' needs, and that users are informed about the quality of statistical outputs, including estimates of the main sources of bias and other errors.'

And finally:

Please keep documented evidence of your quality assurance and value-added efforts described in this paper which may be used by the Statistical Governance Team for submission to the Office for Statistics Regulation as part of their future National Statistics publication assessment exercises. Examples of evidence of quality assurance and trustworthiness are shown in section 3.3 of the Overview of Data Quality of Sources and Outputs[6] document.

[6] http://www.isdscotland.org/About-ISD/About-Our-Statistics/_docs/ISD-Quality-Assurance-Process_v1-0.docx

Appendix 1: Administrative Data Quality Assurance Matrix – Recommendations for NHS National Services Scotland

Level of Assurance	Areas of practice related to quality assurance of administrative data regularly provided for producing Official Statistics			
	Operational context & administrative data collection	**Communication with data supply partners**	**QA principles, standards and checks applied by data suppliers**	**Producer's QA investigations & documentation**
A1 Basic assurance Review and publication of a summary of the administrative data QA arrangements	**Publish** on the publication's accompanying metadata an outline of the administrative data collection processes including the collection stages, steps taken to improve quality, a statement on why quality is important, and any changes to data collection and associated quality implications.	**Documented evidence** of communication with data supplier which contains information about the specification, format, timing (frequency), and sign-off of the data required. **Documented evidence** of communication with data supplier regarding data errors and other quality issues as well as steps to resolve them. **Documented evidence** of views of users (e.g. Scottish Govt) about data quality and resolved quality issues.	**Publish** on the publication's accompanying metadata a brief description of the data suppliers' quality assurance checks. **Documented evidence** of whether supplier carries out internal/operational audits on their admin data and the implications for the statistics.	**Publish** on the publication's accompanying metadata a description of the PHI quality assurance checks, processes and findings for the statistical publication. Include data strengths, limitations and quality risks.
A2 Enhanced assurance Evaluation of the administrative data QA arrangements and publication of a fuller description of the assurance	As above for level A1 plus the following: · A process map showing the collection process · Sources of bias and error in administrative systems or data · Safeguards to minimise data quality risks	As above for level A1 plus the following: · Description of the legal basis for data supply · Data transfer processes · Data protection arrangements · Description of ongoing effective communication mode with data suppliers	As above for level A1 plus the following: · A fuller description of supplier QA checks, principles and indicators within published QA statements · Description of role of IG or information Management groups in QA within documented evidence	As above for level A1 plus the following: · A fuller description of PHI QA checks and indicators including progress against specific QA indicators
A3 Comprehensive assurance Investigation of the administrative data QA arrangements, identification of the results of independent audit, and publication of detailed documentation about the assurance and audit	As above for level A2 plus the following: · Differences across areas in the collection and recording of the data · Commentary on issues with individual data items · Commentary on issues with data completeness, data submission, QA targets and performance · More detailed information on impact of changes in data collection	As above for level A2 plus the following: · A written agreement with data suppliers covering legal basis of data supply, roles and responsibilities, supply and transfer process, security and confidentiality, frequency of data supply, and data specification. · Evidence of a change management process for varying details specified in the agreement · Evidence of regular engagement with users (e.g. by scheduled meetings) to discuss quality issues, specifications, etc.	As above for level A2 plus the following: · Documented evidence of supplier review of QA reports, audits and investigations of received data · Description of why the supplied data continue to be satisfactory for official statistics purposes	As above for level A2 plus the following: · A more detailed description of PHI QA checks, metrics, for specific QA indicators, comparisons with other relevant data sources, effects of QA targets/performance.

**Note that not every recommendation in every cell in the table above necessarily has to be achieved for the corresponding assurance level to be met.*

**It is good practice, when engaging in data quality communication/meetings with data suppliers, to explain that the rationale for continuous engagement with them is 'to continuously improve the quality of ISD's/HPS's statistical outputs in order for our users to continue to trust our statistics'.*