Public Health & Intelligence

NHS National Services Scotland

PHI Analysis Checking Guidance

Document Control	
Version	Version 2.0
Date Issued	September 2016
Author	Richmond Davies
Comments to	NSS.nssstatsgov@nhs.net

Version	Date	Comment	Author
Draft Version 20141001_Checking_v4.docx	09/06/2014	Draft paper developed and shared for discussion with Statistics Advisory Group	Nic Rigglesford
Final Version 20141113_Checking Guidance_final.docx	16/09/2014	Amended following comments from SAG members	Nic Rigglesford
Guidance Paper	03/02/2015	Minor amendments following input from NI&I teams and final SAG discussion. Version created for issue as Guidance	Nic Rigglesford
Version 2.0	01/09/2016	Additions to checking checklist as a result of recommendations following an IG incident	Richmond Davies

Contents

1.0 Introduction

This document provides staff producing analytical outputs with a reference for best practice for checking data analysis. The principles apply to all analytical/statistical work including Publications, project work, information requests (core, non-core and Freedom of Information requests) and parliamentary questions, across all datasets. By following these best practice principles and using the checklist in Appendix 1, you will be making sure that PHI takes a consistent approach to checking all data analysis.

2.0 Before commencing work

2.1 Rerunning previous analysis

If you are replicating a previous analysis (e.g. a routine Publication or repeat request) a large proportion of the analysis is often either a repeat of, or is substantively based on, the previous output. However, you should not simply replicate the analysis – coding rules may have changed, clinical practice may be evolving and previous outputs could contain errors or inconsistencies that haven't previously been detected. Sometimes other sources of information on the topic – either emerging from within PHI or from other organisations – come into the public domain and need to be taken into account or referred to, or it may make sense to adjust our analysis to ensure consistency. (e.g. SG publications on related topics, other UK or international outputs)

2.2 Document the process

If the analysis is likely to be recreated (e.g. for Publications, regular information requests) a standard operating procedure should be created with instructions for producing the required output. This document should include full details of the process, a description of what files are needed or have been created, what they contain and where they have been saved. Documenting the process means that in future, anyone in the team can pick up the specification to continue or repeat the work very easily without trawling through all the files and correspondence.

In all analyses any programs, syntax or queries created to extract and/or manipulate data should be clearly annotated. This should include a record of the author, data source, date created at the start of the program and brief comments throughout the program, where possible, to explain each stage.

2.3 Check codes with clinical coding / public health consultant or other expert if necessary.

Data is often collated using clinical code groupings. Before using these code groupings check them for accuracy and inclusivity by searching relevant dictionaries and other published sources. Clinical input may be required from the CPHM responsible for your service area to understand changes in coding practice at a clinical level over time.

Terminology Services provides support and guidance in the use of clinical classifications, coding and terminology to NHS Scotland and its partners. They are located in Data Management Services and can be contacted on 0131 275 7283 or at NSS.terminologyhelp@nhs.net

2.4 Check Data Completeness

Before releasing or publishing any outputs consideration should be given to how complete the dataset is and whether the analysis gives a robust and complete picture. This should be considered in particular for the most recent time periods available as these may be more incomplete due to late submissions. Care should be taken to examine completeness across the whole dataset and also across subsets of the data. For example, data for Scotland might appear complete enough for analysing but individual NHS Board, hospital or specialty level information may have proportionately larger gaps. Sometimes it may be appropriate to use estimates or imputation to give a more complete picture. However, any decision to do this should first be discussed with and approved by your Service/General Operations Manager and the Head of Profession for Statistics.

Completeness information will be available from the manager/owner of each dataset. Some datasets are managed centrally by Data Management Services; others are managed within teams across PHI. Completeness information for SMR data can be found here: http://www.isdscotland.org/Products-and-Services/Hospital-Records-Data-Monitoring/SMR-Completeness/.

In general terms completeness levels of 95%-98% and over for routine national datasets would be considered robust enough for Publications. Completeness requirements for other types of data (e.g. surveys) and analyses may vary depending on the focus and purpose of the output, and decisions should be made jointly between the dataset manager/owner and the person carrying out the analysis. For example, with developmental or experimental statistics it may be preferable to release outputs based on lower completeness levels to publicise the work and focus attention on the developing data. If further advice is required, speak to your Service/General Operations Manager or the Head of Profession for Statistics.

2.5 Check Data Quality Issues

Even if the dataset appears complete, there may be problems or issues with the content of fields within it. Before beginning to analyse the dataset you should investigate patterns in the data and reassure yourself that the data appear to have been recorded accurately enough for the requirements of your output. For example, it may be appropriate to look at long term trends in key fields, frequencies, comparisons of coding between different Boards, coding over time etc.

In addition the manager/owner of the dataset should have information about known issues – e.g. changes in definitions over time, closure of facilities. There is a Metadata Repository pulling together information about datasets and data issues contained in ACaDMe and SMR data sets. The intention is that this repository will be extended to include metadata on other datamarts and datasets:

https://metadata.nhsnss.scot.nhs.uk/

Again, some datasets are managed centrally by Data Management Services; others are managed within teams across PHI. The National Data Catalogue contains a full list of data sets that ISD holds along with summary information and contact details:

http://www.ndc.scot.nhs.uk/National-Datasets/Full-A-Z/index.asp

If the dataset you are analysing has a User Group (e.g. the Scottish Morbidity Record (SMR) Analysts Forum covers all SMR datasets) you should contact a member of this group for details about known issues.

The dataset manager/owner and the person carrying out the analysis have joint responsibility for investigating and understanding both completeness and quality. The person carrying out the analysis should then make the decision on what is appropriate for release to a customer or reporting within a Publication, including which time periods are suitable for release.

Care should be taken to ensure that information requests, media queries and PQs are only answered using data up to the time period that is available in any related Official Statistics publication. If you need clarification you can get advice via the Head of Profession (Statistics Support Team (nss.phiHOP@nhs.net)).

3.0 Output checks

Staff who are working with any dataset should spend time 'getting to know' the data they work with. Familiarity with expected codes, frequencies and patterns in the data should underpin all analytical work carried out across PHI.

Quality assurance checks should be performed throughout the analytical process to ensure that the output is accurate. These should be documented in the Standard Operating Procedure if the analysis is run regularly, or documented in programmes for individual pieces of work.

In addition, a thorough set of checks should be performed on the final output. The best practice principles below assist both with interim checking as well as the final checking but should not be taken as a complete list – other checks may also be required for individual datasets and/or outputs.

The final checking of the analysis is a two-person process with the first stage of checking carried out by the person who performed the analysis.

The second person check is only intended to be a 'sanity check' – the second person should not have to re-run the entire analysis.

Best Practice for checking : First person check (the person who carries out the analysis)

3.1 Check syntax / method (files/selection criteria/codes)

Check the program is doing what you want it to do.
Check that the data are being read in accurately – for example are positions of variables correct?
Are you using the most up to date file layouts?
Are you selecting the correct codes?
Have you made all the selections required?

3.2 Sense checks of figures

Check that figures are what would be expected, that is they are not too high or low. It's not always possible to know exactly what figures to expect although a general check that figures seem sensible should be done. For example, if analysis is presented by NHS Board it might be reasonable to assume that larger population Boards, such as Greater Glasgow, would have higher figures than a small Board like Orkney.

Check whether sub-totals add up to overall totals

Check whether any of the figures and/or geographies are small and therefore require suppression in line with Disclosure principles.

Outliers or unusual patterns should generally be considered to be a data quality issue until thorough investigation has eliminated this possibility. The manager/owner of the dataset will be able to advise on known issues (see 'Checking Data Quality issues' above); it will sometimes be necessary to contact the data supplier with specific queries, the data manager/owner should advise on the best way to do this.

3.3 Compare analysis with other sources / previous analysis / publications

Look at other publications where possible (PHI or external) for previous work done on the same topic to make sure the results are similar. For example, National Records of Scotland (NRS) publish extensive information on deaths that can be used to check related PHI outputs against.

Data may have to be aggregated up to find comparable source, for example if analysis is at postcode sector level, this information is not likely to be available on the web, although data at NHS Board or Scotland level may be. Be careful to make sure that the data you are checking against are comparable with the analysis, for example same coding selections made.

3.4 Check by rerunning analysis in another package

If there are other ways of running the analysis, it may be worthwhile to redo 'headline numbers' for the job in another format.

Or alternatively, where possible, you could run your original program against a test data set where the results are known.

3.5 Check transfer of figures from output to a different package

If the output is transferred or copied from statistical software to another package, such as Excel, prior to release check the output in the final table for any mistakes made whilst transferring the data. This is a common source of mistakes. It is a good idea to temporarily save any extract or output files created during the analysis so that if a mistake is found, it can be pinpointed to a mistake from running the analysis or creating the final tables.

If any calculations are carried out after transferring the output, check that they are correct – for example, manually check some of the calculated figures.

In particular, beware of the following: formulae that have been 'dragged down'; data that is in a different order within the data output from the order of row/column headers in the excel spreadsheet being copied into (for example, NHS Board alphabetised by cipher vs. NHS Board alphabetised by description); data that has been 'sorted' without selecting the entire array of data.

3.6 Format checks on final Table/Chart.

Housekeeping
Check that titles are correct, sources are added and that all relevant notes / exclusions / caveats are mentioned.

Spelling check

Spell-check the final output to get rid of any typing errors.

Overall presentation – is it clear?
Consider whether the final tables and report are easy to interpret; and that the titles, column and row headers are easily understood.

The Web and Publications Teams can advise on Accessibility Guidelines which are a statutory requirement for published outputs.

Data visualisations
Any data visualisations should be discussed with a member of the TrIP Data Visualisation Core Group.

Page set up
Check print preview of excel tables to make sure the page settings are appropriate. Spreadsheets should be saved so that they open appropriately - at cell A1 on the first sheet or on an index page for example.

Formats
All formatting should be consistent throughout the documents and tables.

Best practice for checking: Second person check (not the person who carried out the original analysis)

The role of the second person check is a "sanity check". The second person should double check the following aspects of the first person check:

1. Match final output to specification of request.
2. Check syntax/method
3. Sense check of figures
4. Compare analysis with other sources / previous analysis / publications
5. Check by rerunning high level analysis in another package
6. Check transfer of figures from output to a different package
7. Format checks on final Table/Chart.
8. Check supporting correspondence to customer.

The range and depth of checks carried out by the second person will be determined by the particular analysis being produced, local team protocols and your Service/Group Operations Manager. However the second person check should not be as time-consuming as the original running of the analysis.

The second person should note that even though the analyst may have completed the analysis in two different packages as a way of checking, they could have made the same mistakes in each package, for example selected the incorrect code range. Therefore programs/syntax should be checked carefully.

On occasion (and for all Publication outputs) Service/Group Operations Managers will perform a final sign off check. A record of the checking undertaken should be incorporated into the SOP.

4.0 Exceptions – turnaround times

On rare occasions, in particular during public health incidents, teams across PHI have to react rapidly and turn outputs around very quickly. The overarching principles set out above still apply in these circumstances, however, the detailed checks may not be possible. In these circumstances ongoing knowledge and expertise in the dataset, as well as routine quality assurance processes, should allow for confidence in outputs.

Appendix 1: Summary Checklists

First person check

Before you start your analysis you should:

- £ Know/find out who is responsible for the dataset(s) you are analysing (the 'owner')
- £ Check that you have read and understood the details of any relevant Confidential data Release Form (if applicable)
- £ Check that you are familiar with what is in and out of scope as detailed in any Public Benefit and Privacy Panel (PBPP) authorisation (if applicable)
- £ Check whether the same/similar analysis has been produced before
- £ Decide on required documentation (commented program, full SOP etc)
- £ Discuss coding choices with the relevant clinical expert.
- £ Check for data completeness issues
- £ Check for data quality issues, referring to established metadata where available
- £ Check which is the most recent time period published in any Official Statistics release and therefore available for use for information request/PQ responses

When checking your outputs you should:

A: check your syntax and data selections

- £ Check the programme/query is doing what you expected
- £ Check data are being read in accurately
- £ Check you are using the latest file layouts
- £ Check you are selecting the correct codes
- £ Check you have selected all the required fields/inclusions/exclusions

B: perform basic logic checks of the output

- £ Check that figures are what you would expect
- £ Check that comparisons between the Boards look reasonable

Remember – assume outliers or unusual patterns should be considered to be a data quality issue until thorough investigation has eliminated this possibility

- £ Check that subtotals add up to totals
- £ Check if suppression is required for Disclosure Control purposes

C: compare your outputs with other sources

- £ Check your outputs against other publications/outputs (PHI or external) that contain the same or related information.

D: rerun the analysis

- £ Check that an alternate analytical package gives the same result
  Or
- £ Check that the same program/analysis gives accurate results on a set of known data (e.g. previous year)

E: check data transfers

- £ Check that data has been transferred accurately between software packages (e.g. from BOXI to excel)
- £ Check formulae are accurate (e.g. in excel calculated cells)
- £ Check formulae that have been 'dragged' are accurate for all cells
- £ Check that NHS Board descriptions and data are in the same order
- £ Check that any 'sorts' have included the appropriate data array
- £ Check that figures quoted in narrative sections are accurate

F: formatting and presentation

- £ Check all titles are correct
- £ Check all data sources are listed
- £ Check all caveats and footnotes are added and accurate
- £ Check spelling
- £ Check for overall understanding and clarity of the output
- £ Check any data visualisations with the Data Visualisation Core Group
- £ Check for accessibility
- £ Check print previews to ensure page settings are appropriate
- £ Check excel sheets open on the required sheet, curser on cell A1
- £ Check formatting is consistent across all tables/documents/charts

After analysis is complete you should:

- £ Document the analysis carried out/comment within programmes/syntax
- £ Update Standard Operating Procedures if required
- £ Confirm that the fields included in the final output file match the specification and any confidential data release forms / Public Benefit and Privacy Panel approvals
- £ Ensure that we have permission to provide reference identifiers, including CHI number
- £ Ensure that, for anonymised extracts, identifiable fields such as CHI have not been included
- £ Carry out statistical disclosure control assessment (contact nss.nssstatsgov@nhs.net for advice if needed)

Second person check

The range and depth of checks carried out by the second person will be determined by the particular analysis being produced, local team protocols and your Service/Group Operations Manager but you should consider:

- £ Check the final output matches the specification of request.
- £ Check syntax/method
- £ Sense check of figures
- £ Check analysis against other sources / previous analysis / publications
- £ Check by rerunning high level analysis in another package
- £ Check transfer of figures from output to a different package
- £ Format checks on final Table/Chart.

- £ Check supporting correspondence to customer.
- £ Confirm that the fields included in the final output file match the specification and any confidential data release forms / Public Benefit and Privacy Panel approvals
- £ Ensure that we have permission to provide reference identifiers, including CHI number
- £ Ensure that, for anonymised extracts, identifiable fields such as CHI have not been included
- £ Check that statistical disclosure assessment has been carried out