ESSNet S D C

A Network of Excellence in the European Statistical System in the field of Statistical Disclosure Control

Guidelines for the checking of output based on microdata research

Maurice Brandt (Destatis), Luisa Franconi (ISTAT), Christopher Guerke (Destatis), Anco Hundepool (CBS), Maurizio Lucarelli (ISTAT), Jan Mol (CBS), Felix Ritchie (ONS), Giovanni Seri (ISTAT), Richard Welpton (ONS)[1]

Summary: In this document practical guidelines for output checking are given. Output checking is the process of checking the disclosure risk of research results based on microdata files made available in Research Data Centers. Actual rules for different types of output are given. These rules are placed within a framework of two different models for output checking. Also, recommendations and best practices for some organisational and procedural aspects of the output checking process are given.

Keywords: Research Data Center, disclosure control, output checking, guidelines

[1] This document would not have looked as nice without the efforts of Tanvi Desai at ONS for proofreading and editing it. Any remaining errors are of course solely ours.

Table of contents

1. Introduction and general approach

1.1 Introduction

Most National Statistical Institutes (NSIs) realise that the full potential of their microdata can never be extracted by just their own staff and in their own official publications. The shift from survey-based to register-based statistics has led to a rapid increase in the amount of available microdata. At the same time, IT developments have made it possible to analyse these very large datasets, but a lot of NSIs simply do not have the resources available to make full use of these available datasets. Luckily, a very large number of qualified researchers in the world of academia and policy-making are willing to share in this work. NSIs therefore provide researchers access to their microdata files, for instance in a Research Data Center (RDC). NSIs must then find the right balance between enabling these researchers to do their work while always ensuring the confidentiality of the data.

One common method of protecting the confidentiality of the data is to check all results that researchers want to take out of the controlled environment of the RDC. Throughout this document, this process will be referred to as *output checking*.

This document provides the reader with guidelines for, and best practices of this output checking process. The aim of the document is threefold:

1. to allow experienced NSIs to learn from each other by sharing best practice
2. to provide NSIs that have little or no experience in RDCs with practical advice on how to set up an efficient and safe output checking process.
3. to facilitate harmonisation of output checking methods across NSIs.

The third aim is particularly crucial in light of the current focus on the development of European infrastructures for microdata access, as it is crucial that NSIs agree on a common method for checking output. As cross-border data access will be most efficient if NSI X can check output that was generated with datasets from country Y as well. Only with common guidelines will NSIs delegate the task of output checking to each other.

The remainder of this chapter deals with a general approach to output checking. In chapter 2, possible types of output will be categorized. For each category of output both a simple rule and guidance for more experienced researchers will be outlined. Chapter 3 then deals with a number of practical issues (procedural, organisational), which are necessary for an efficient high-quality output checking process. Chapter 4 focuses on the training for researchers that use the RDC and will be producing the output.

1.2 General approach

Before continuing with the rest of the document we first need to discuss the general approach to output checking.

The RDC zoo

An RDC is a safe environment were accredited researchers can access the most detailed micro data to undertake any research that they desire (as long as it serves the public good). This makes output checking in an RDC totally different from disclosure control of official NSI publications. The official publications are of a well defined form (usually a table) and the intruder scenarios are limited in number. Whereas the output of an RDC can be anything! Researchers twist, transform and link the original data in different and complex new ways. This makes it very difficult to come up with a set of rules for that cover every possible output, as one expert vividly states it: designing rules for output checking is like designing cages for a zoo - that will keep the animals both contained *and alive* - without knowing in advance which animals will be kept in the cages[2].

Safe and unsafe classes of output[3]

To bring some order to output checking in an RDC all output can be classified into a limited number of categories (for instance tables, regression etc.). Each class of output is then labelled 'safe' or 'unsafe'. This classification is done solely on the functional form of the output, not on the data itself.

- 'safe' outputs are those which the researcher should expect to have cleared with no or minimal further changes; for examples, the coefficients estimated from a survival analysis. Analytical outputs and estimated parameters are usually 'safe'. The exceptions where a 'safe' output is not released should be well defined and limited in number.
- 'unsafe' outputs will not be cleared unless the researcher can demonstrate, to the output checker's satisfaction, that the particular context and content of the output makes it non-disclosive. For example, a table will not be released unless it can be demonstrated that there are enough observations, or the data has been transformed enough, so that the publication of that table would not lead to identification of outputs. Linear aggregations of data are almost always 'unsafe'.

---

[2] Ritchie, F (2007) *Statistical Disclosure Control in a Research Environment*. Mimeo: Office for National Statistics

[3] For a more detailed explanation, see Ritchie, F. (2008) "Disclosure detection in research environments in practice", in Work session on statistical data confidentiality 2007; Eurostat; pp399-406

Note that the burden of proof differs: for safe outputs, the output checker must provide reasons why the output cannot be released against normal expectations; for unsafe outputs, the researcher has to make a case for release. In both situations, the ultimate decision to release an output remains with the output checker.

Output checking is always context specific. It is not possible to say, ex ante, that something will or will not be released. The purpose of the safe/unsafe classification is to give guidelines on the likelihood of an output being released – and, if it is unsafe, to suggest ways that it can be made safe.

Not all outputs have yet been classified. The default classification is 'unsafe'.

Two types of error

With this in mind, one needs to realise that the optimal way to check output, is the one that maximises use of the datasets while minimising the risk for disclosure. So, phrased differently, a less than perfect output check can lead to two types of errors:

1. Confidentiality errors: releasing disclosive output
2. Inefficiency errors: not releasing safe output

Rules and guidelines for output checking can prevent both errors. But the trick is to find the right rule.

Consider, for instance, a rule that sets a minimum for the number of units in each cell of a table (threshold rule). If the minimum cell count is set too high, this will lead to inefficiency errors: the output will be safe, but will contain less information than could be obtained from the datafile. The researcher might have had to group classes that he was interested in, to reach the threshold. On the other hand, if the minimum is set too low, this will lead to unsafe outputs being released.

Finding the correct disclosure control rules for use in an RDC is especially difficult. One has to realise that the exact shape of the output is not known beforehand. The idea of an RDC is to give researchers the maximum amount of freedom in analysing the data files, so they will produce output of all sizes and shapes (tables, models etc).

To deal with this problem, a two-model approach has been developed. The first model is called the principles-based model. This model minimises both confidentiality and inefficiency errors. The other is called the rule-of-thumb model. For this model, the focus is on preventing confidentiality errors and inefficiency errors are accepted. Both models will be discussed in more detail in the following paragraphs.

1.3 Principles-based model

The principles-based model centres on a good collaboration between researchers and RDC staff. Because this model also aims to prevent inefficiency errors, simple rules for checking output are not appropriate. The reason is that simple rules can never take into account the full complexity of research output. To give maximum flexibility to the researcher, no output is ruled in or out in advance. All output needs to be considered in its entire context before deciding on its safety. For instance, a table that contains very small cell counts (maybe even some cell counts of 1) isn't necessarily unsafe. If, for instance, the original data was transformed beforehand, the information that the 'risky' cells disclose, might not be traceable to the individual. What is needed is a clear understanding of the governing principles behind disclosure control, therefore, both the researcher and the RDC staff need training in disclosure control.

The principles-based model has the obvious advantage that it leaves a maximum amount of flexibility to the researcher. Data files will therefore be used to their fullest extent. However, the model also has some possible drawbacks:

- The model relies on serious training of NSI staff and researchers. Researchers have to be willing to invest their time and effort on a topic, which is not naturally within their field of interest.
- The model spreads the responsibility for clearing an output. In a rules-based model, the responsibility lies with the people that design the rules. In this principles-based model, the responsibility lies with each individual checker. There are no strict rules to follow and each checker has to make his own decision on clearing the output, based on his experience and understanding of the underlying principles.

To circumvent these drawbacks, an alternative model is presented: the rule-of-thumb model.

1.4 Rule-of-thumb model

In this model the main focus is on preventing confidentiality errors, and some inefficiency errors are taken for granted. This model typically leads to very strict rules. The chance that an output, that passes these rules, is non-disclosive is very high. The advantage is that the rules can be applied more or less automatically by both researchers and staff members with only limited knowledge of disclosure control.

It is important to stress the fact that, although the rules of thumb are very strict, this is not a 100 % guarantee that all output that passes these rules is indeed non-disclosive. There is a very small chance that a disclosive output slips through. This is because the rules are rigid and do not take the full context of the output into consideration.

The rule-of-thumb model is useful for a number of situations:

- Naïve researchers whose output is usually far from the cutting edge of disclosure control (for instance policy makers who just want tabular output with limited detail)
- Inexperienced NSIs starting up an RDC. In this case, both users and RDC staff could have too little experience to be able to work with the principles-based model. The rule-of-thumb model provides them with a starting point that ensures maximum safety. In using the rule-of-thumb model, they build up experience along the way. At some point in time they might feel confident enough to set up the principles-based model and open up the way to clearing more complex output.
- Automatic disclosure control for RDCs. This will mainly be useful for more controlled types of data access like remote execution. In remote execution, researchers write their scripts without having access to the real datafile (sometimes dummy datasets are provided for this purpose). They then send the finished script to the RDC, where a staff member (or an automated system) runs it on the full datasets. The results are then returned to the researcher.

Even for RDCs using the principles-based model, the rules of thumb are usually the starting point when checking any particular output. Using the rules of thumb, attention is quickly focused on the parts of the output that breach these rules. These parts can then be considered more carefully using the full principles-based model to decide whether they can be released or not.

2. Rules for output checking

2.1 Classification of output

As described before, the RDC zoo can be somewhat structured by classifying all output into a limited number of classes. The table below lists the different classes of output. Each class is marked as either safe or unsafe (see section 1.2 for an explanation of the safe-unsafe classification).

Type of Statistics	Type of Output	Classification
Descriptive statistics	Frequency tables	Unsafe
	Magnitude tables	Unsafe
	Maxima, minima and percentiles (incl. median)	Unsafe
	Mode	Safe
	Means, indices, ratios, indicators	Unsafe
	Concentration ratios	Safe
	Higher moments of distributions (incl. variance, covariance, kurtosis, skewness)	Safe
	Graphs: pictorial representations of actual data	Unsafe
Correlation and Regression Analysis	Linear regression coefficients	Safe
	Non-linear regression coefficients	Safe
	Estimation residuals	Unsafe
	Summary and test statistics from estimates (R^2, χ^2 etc)	Safe
	Correlation coefficients	Safe

2.2 The overall rule of thumb

As discussed before, the rule-of-thumb model is based on clear and simple (and strict) rules. Because these rules differ only slightly for different classes of output, an overall rule of thumb can be established. This overall rule of thumb is presented first; after that, when describing the rules for each class of output, the interpretation of this overall rule of thumb, for the specific class of output is given.

The overall rule of thumb has four parts:

1. **10 units**: all tabular and similar output should have at least 10 units (unweighted) underlying any cell or data point presented. A common term for such a rule is a threshold rule (the cell count must exceed a specified threshold).

2. **10 degrees of freedom**: all modelled output should have at least 10 degrees of freedom and at least 10 units have been used to produce the model. Degrees of freedom = (number of observations) -/- (number of parameters) -/- (other restrictions of the model)
3. **Group disclosure**: in all tabular and similar output no cell can contain more than 90 % of the total number of units in its row or column to prevent group disclosure. Group disclosure is the situation where some variables in a table (usually spanning variables) define a group of units and other variables in the table divulge information that is valid for each member of the group. Even though no individual unit can be recognized, confidentiality is breached because the information is valid for each member of the group and the group as such is recognizable.
4. **Dominance**: in all tabular and similar data the largest contributor of a cell can not exceed 50 % of the cell total.

A practical problem: The dominance rule

As simple as it looks, even the overall rule of thumb has one major difficulty. This concerns element n^{o} 4: the dominance rule.

In order to check this rule, the researcher must provide additional information on the value of the largest contributor for each cell. This often burdens the researcher with a lot of extra work. In addition the extra information obviously has to be removed from the output before release, as releasing the value of the largest contributor of a cell would be very disclosive!

So, although the dominance rule is included in the rule of thumb, the current practice in many countries is that it is not actively checked. Even so, researchers are told that they should take it into consideration when creating their output.

Usually, an NSI decides to actively check it only in certain circumstances, for instance:

- magnitude tables on business data
- output based on very sensitive variables
- variables with a very skewed distribution.

Nevertheless, for those wishing to follow only the rules of thumb rather than the principles-based model, checking for dominance should be considered best practice.

The remainder of this chapter will examine each output classified above and discuss the interpretation of the overall rules of thumb and give some more detailed information for the principles-based model.

2.3 Class 1: Frequency tables

2.3.1 Rule of thumb

In the case of frequency tables only parts 1 and 3 of the overall rule of thumb apply:

- Each cell of the table must contain at least 10 units (unweighted). Researchers should include the unweighted cell count in their tables to enable the checking of this rule.
- The distribution of units over all cells in any row or column is such than no cell contains more than 90 % of the total number of units in that particular row or column (no concentration of units in only one cell of the row or column). This rule prevents group disclosure.

2.3.2 *Detailed information for principles-based mode*

Every cell in a table is potentially unsafe. There are no general rules for making a table safe. However, as noted above, there are various options for making a table safer. In such circumstances, a number of issues should be taken into consideration:

- whether the data is itself disclosive (whether it has been transformed into a new variable; level of detail, etc)
- whether units making up the data or subsets could be identified
- closeness of the data to elements 1 (threshold), 3 (group disclosure) and 4 (dominance) of the rule of thumb
- whether the rank ordering of contributors is known (in other words, is the largest/smallest/tallest etc known or guessable?)
- choice of the cell units; are these people, households, regions etc
- sample choice
- weighting

The context in which analysis is undertaken is important. Factors that should be taken into consideration include:

- level of geographic disaggregation
- detail of industrial/occupational classification
- global context: domestic vs. international operations

The most important criteria are establishing that no respondent can reasonably be identified, and that no previously unavailable confidential information about groups can be inferred.

'Reasonably' is not explicitly defined – all definitions are subject to criticism.; moreover, the aim of principles-based output checking is that all definitions are subject to the particular context. However, some factors which would be taken into consideration include identification

- taking a significant amount of time and effort
- requiring additional knowledge which most individuals would not be expected to have or be able to acquire easily
- needing some technical ability

2.4 Class 2: Magnitude tables

2.4.1 Rule of thumb

In the case of magnitude tables elements 1, 3 and 4 of the overall rule of thumb apply:

- Each cell of the table contains at least 10 units (unweighted). Researchers should include the unweighted cell count in their tables to enable the checking of this rule.
- The distribution of units over all cells in any row or column is such than no cell contains more than 90 % of the total number of units in that particular row or column (no concentration of units in only one cell of the row or column). This rule prevents group disclosure.
- In every cell, the largest contributor can not exceed 50 % of the cell total.

2.4.2 Detailed information for principles-based mode

There may be circumstances in which the cell count rule of ten is inappropriate. For example, a researcher may believe that an output is non-disclosive even though cell counts do not meet the required threshold of ten units. The onus is then on the researcher to explain why this is the case, and the individual responsible for checking output will be required to make a decision as to whether this output can be released.

The following factors are important in making this decision:

- the context of the output
- accompanying information (this may be dependent on previous outputs)
- can the person checking output identify individuals or companies from the output, particularly if the output includes a dominant observation?

The guidelines in section 2.3.2 should be borne in mind when making this judgement.

2.5 Class 3: maxima, minima, percentiles

2.5.1 Rule of thumb

Maxima and minima are not released since they usually refer to only one unit.

Percentiles are treated as special cases of magnitude tables. Each percentile band should be treated as a table cell with membership of the cell determined by position in the rank. If a unit's position is known, then information about that unit can be gleaned. How useful/disclosive this is depends upon the size of the cell and the range of the band. The rules of thumb for magnitude tables apply.

2.5.2 Detailed information for principles-based model

For percentiles, the same principles as for magnitude tables apply.

In the principles-based model, for maxima and minima the rules for magnitude tables also apply. This means that if the minimum/maximum can not be associated with an individual data point, they can be released.

2.6 Class 4: Modes

2.6.1 Rule of thumb

The mode is safe when element 3 of the rule of thumb (group disclosure) is met. This prevents the case that all observations have the same value.

2.6.2 Detailed information for principles-based model

The rule of thumb applies.

2.7 Class 5: Means, indices, ratios, indicators

2.7.1 Rule of thumb

Common SDC rules can be applied to means, indices, ratios and indicators, as each represents the synthesis of *n* observed values in a single value. The same considerations as for magnitude tables cells apply; in particular, elements 1 and 3 of the overall rule of thumb apply:

- Each single value should derive from the synthesis of at least 10 units (unweighted). Researchers should include information in their output to enable the checking of this rule.
- For each single value to be released, the largest contributor included in the synthesis can not exceed 50 % of the total. Researchers should include information in their output to enable the checking of this rule.

2.7.2 Detailed information for principles-based model

When evaluating means, indices, ratios and indicators, the following considerations should be taken into account.

First of all, index formula should be considered.

In general, a simple index summarizes the individual variable values for the statistical units in a given population:

$$I = f(X, n)$$

For the output evaluation, the index formula (*f*) and the population size (*n*) must be specified.

Sometimes the index formula is very complex, involving a lot of attributes for each unit and combining values in such a way that reverse calculation of single values is unrealistic.

As an example, consider the complexity of the Fisher price Index:

$$P_F = \sqrt{P_L \cdot P_P} = \sqrt{\frac{\sum_{j=1}^{m} p_{1,j} \cdot q_{0,j}}{\sum_{j=1}^{m} p_{0,j} \cdot q_{0,j}} \cdot \frac{\sum_{j=1}^{m} p_{1,j} \cdot q_{1,j}}{\sum_{j=1}^{m} p_{0,j} \cdot q_{1,J}}}$$

Therefore, assuming that the index is not calculated upon very few units, complexity of data transformation is itself a reasonable protection against the disclosure of individual information from the value of the index. In the same way, complex indices (i.e. a combination of several simple indices, including ratios) are in general less disclosive than simple ones.

Nevertheless, index formula can also be very simple, e.g.:

a) $I = \frac{\sum_{i=1}^{n} X_i}{n}$, b) $I = \frac{\sum_{i=1}^{n} X_i}{\sum_{i=1}^{N} Y_i}$, c) $Range = X_{MAX} - X_{MIN}$

Furthermore, the value of one or more arguments of the index formula could be easily publicly available. Note that a) corresponds to the arithmetic mean.

For instance, in case a) and b) the denominator could be known: in such cases, the problem is reduced to the evaluation of the numerator ($\sum X_i$). This matches the issue of checking a cell of a magnitude table (if X is quantitative) or a frequency table (if X is dichotomous).

Only in this last particular case, where X={0,1}, coherently with the case of frequency tables, the applied rule should be: $\sum X_i \geq 10$, and, since the frequency count of '0' values can also be derived, also: n-$\sum X_i \geq 10$.

Example: with X dichotomous, a mean value $\overline{X} = 0.7$ with a population size of n=20 corresponds to the frequency table cells:

X=0	X=1	Total
6	14	20

In this example, even if the mean of X is calculated upon 20 units (more than 10) and the corresponding frequency of values '1' is 14 > 10 (rule for frequency cell counts), one can derive that there are only 6 (<10!) units having X=0. However, it has to be considered whether, in this kind of situation, there is an effective risk of disclosing information upon individuals or not (see the section on frequency tables 2.3).

About case c), being Range a linear function of maximum and minimum, it should be released only if its components (max and min) could be released (see section 2.5). In the same way, as stated before, when evaluating other comparison or dispersion indices, the formula (whether it is linear or not) and its arguments should be considered.

Nevertheless, also publicity of involved variables should be considered.

The 'Indices' category comprises a wide range of statistical results (including means, totals, etc.). From the output checking point of view indices pose an SDC problem practically quite easy to be dealt with, but theoretically analogous to that of tabular data. Therefore, the same items have to be taken into account (see section 2.3 and 2.4), particularly the type of variables involved (if they are publicly available as often is the case for social indices) and the characteristics of the sub-population of n units involved in the index computation (if they are selected according to sensitive or identifying spanning variables).

2.8 Class 6: Concentration ratios

2.8.1 Rule of thumb

Concentration ratios are safe as long as they meet elements 1 (threshold), 3 (group disclosure) and 4 (dominance) of the rule of thumb.

2.8.2 Detailed information for principles-based model

Concentration ratios below CR10 (largest ten units) are allowable provided the results can be shown to be non-disclosive, which would be the case if

- the sample on which the output is based contains at least ten units
- rule 4: the dominance rule is met
- percentages are displayed with no decimal places

2.9 Class 7: Higher moments of distributions

2.9.1 Rule of thumb

Higher moments (variance, skewness, kurtosis) are safe as long as part 2 of the rule of thumb is met:

- all modelled output should have at least 10 degrees of freedom and at least 10 units have been used to produce the model.
  Degrees of freedom = (number of observations) -/- (number of parameters) -/- (other restrictions of the model)

2.9.2 Detailed information for principles-based model

The simple rule applies. In discussions with researchers, it may be noted that with less then ten degrees of freedom the statistical value of the output is doubtful anyway.

2.10 Class 8: Graphs (descriptive statistics or fitted values)

2.10.1 Rule of thumb

Graphs in themselves are not permitted as output. The underlying data may be submitted as output, which when cleared, may be used to reconstruct graphs in the researcher's own environment.

2.10.2 Detailed information for principles-based model

Graphical output is only allowed in cases where it is impossible to reproduce the graph without access to the full microdata; in other words, to build the graph from (tabular) output that will be cleared. In this case the graph should meet the following conditions:

- Data points can not be identified with units. When the graph consists of transformed or fitted data, this is usually not a problem.
- There are no significant outliers
- The graph is submitted as a 'fixed' picture, with no data attached. This means that graphs should be submitted as files with one of the following extensions: .jpg - .jpeg - .bmp - .wmf

2.11 Class 9: Linear regression coefficients

2.11.1 Rule of thumb

Linear regression coefficients are safe provided at least one estimated coefficient is withheld (e.g. intercept).

2.11.2 Detailed information for principles-based model

Complete regressions can be released as long as

- they have at least 10 degrees of freedom
- they are not based solely on categorical variables
- they are not on one unit (e.g. time series on one company)

2.12 Class 10: Non-linear regression coefficients

2.12.1 Rule of thumb

As with linear regressions, non-linear regression coefficients are safe provided at least one estimated coefficient is withheld (e.g. intercept).

2.12.2 Detailed information for principles-based model

Non-linear regressions differ from linear regressions because of the nature of the dependent variables, which are discrete. If a regression is estimated, and the same regression is repeated with one additional observation, then by using the variables means in conjunction with the regression results, it may be possible to infer information about that one particular observation.

However, in practice, this is unlikely. Changes in the number of observations included in a model are the result of

- changing the sample explicitly, or
- changing the specification and so finding observations with inadmissible values (eg missing) dropping out

The first is unlikely to lead to one observation more or less, as the information gain is negligible (unless the researcher is deliberately aiming to circumvent SDC rules). In the second case, different numbers of observations are not relevant because the specification has been changed

2.13 Class 11: Estimation residuals

2.13.1 Rule of thumb

No residuals and no plots of residuals should be released.

2.13.2 Detailed information for principles-based model

As with the rule of thumb, residuals should not be released.

A reasonable request for a plot of residuals may be made by a researcher, for example, in order to demonstrate the robustness of the model. However, plots of residuals should be discouraged. Instead, a research should analyse plots within the safe setting, and a written description of the shape of the plot may be released to demonstrate robustness. If there is a need for a plot of residuals to be released, this should be assessed as for graphs (2.10) above.

2.14 Class 12: Summary and test statistics

2.14.1 Rule of thumb

The following summary statistics can be released provided there are at least 10 degrees of freedom in the model:

- R2 and variations
- estimated variance
- information criteria (e.g. AIC, BIC)
- individual and group tests and statistics (t, F, chi-square, Wald, Hausman etc)

2.14.2 Detailed information for principles-based model

No principles in addition to the rule-of-thumb guidelines.

2.15 Class 13: Correlation coefficients

2.15.1 Rule of thumb

Correlation is a measure of linear relationships between variables. The checkers have to ensure that the released outputs cover first element of the rule-of-thumb, which means a minimum of 10 unweighted units underlying each correlation coefficient.

2.15.2 Detailed information for principles-based model

A very small number of cases exist where problems could arise (e.g. the correlation matrix includes 0 or 1). Even in these cases, the problem is the publication of correlation coefficient connected with summary statistics.

Correlations of binary variables are treated like linear regressions with binary explanatory variables or full saturated linear regressions. Therefore, the same items have to be taken account as for linear regression coefficients (see section 2.11). Nevertheless, in the majority of analyses, correlation coefficients could be classified as safe.

3. Organisational/procedural aspects of output checking

3.1 General remarks

The previous chapter dealt with rules for output checking. An efficient, high-quality output checking process is only possible when effective organisational and procedural practices are implemented in addition to these rules. In this chapter the most important of these will be discussed. Practical guidelines are given for each, split into a minimum requirement and a best practice. The difference between the minimum requirement and the best practice is often just a different balance between cost and benefit. The minimum can be seen as the best value-for-money measure, while the best practice is the operational excellence every RDC should aim for.

3.2 Legal basis

The aim of this section is to outline the legal framework for output checking. Output (and therefore of output checking) is essential to an RDC. Because an RDC provides a researcher access to highly confidential data, some of the NSIs responsibility for confidentiality needs to be transferred to the researcher. This responsibility can be underpinned by drawing up contracts that include the dos and don'ts for researchers using an RDC.

Minimum standard

A legally enforceable agreement with the researcher needs to be drawn up. As a minimum standard this contract should include the statement that data on individual persons/households/companies/organisations etc can NEVER leave the safe environment of the RDC, and that the researcher has a responsibility to ensure this.

Best practice

As a best practice a two-tier approach is suggested. The first level is a contract with the research organisation that the user is affiliated with. This contract makes the organisation as a whole responsible for the research project and for keeping the data safe. The agreement contains confidentiality statements, rules for an appropriate use of the data (what can/cannot be done with the data, including rules relating to NSI policy e.g., not reproducing official statistics, not duplicating other research, etc.), and details of penalties for misuse of the data. This agreement ensures that if employee A of this organisation breaches confidentiality, it can have repercussions not only for employee A, but for the organisation as a whole. Given the seriousness of the breach, the NSI can then decide to ban the whole organisation and not just employer A.

The second level is a confidentiality statement signed by each individual researcher, binding him to safeguard data confidentiality. To link these statements to the institutional contracts, the confidentiality statement should also be signed by the organisation). This means by someone with institutional authority and the ability to discipline employers that misbehave, and not just a senior researcher or professor.

Of course these agreements and statements have to be signed by the NSI as well. The person who signs on behalf of the NSI strongly depends on the national organisation. Four different situations have been identified:

- the RDC director: the person in charge of the RDC has the power to give permission;
- the NSI/Institute: President/Director/Legal unit/Board of Directors;
- the Director of the department responsible for the data;
- a Statistical/Confidentiality Committee: this can be internal or external to the NSI/institution.

The first two situations are the most frequently observed.

3.3 Access requests

An access request is a key component of the organisational and procedural aspect of running an RDC. At the access request stage the following points should be made clear: (i) who is applying to access the RDC (is he/she or his/her institute eligible?); (ii) which results are going to be produced (are they feasible?); (iii) how results will be disseminated (is this consistent with the NSIs policy?). Even where these issues do not directly affect output checking, an accurate preliminarily evaluation of the access request can make the checker's job easier (see for instance section 3.6).

Minimum standard

Access is granted to 'researchers' for 'research purposes': RDCs must deal with qualified and eligible users. Eligibility is defined depending on national/European legislation or on admissibility criteria defined by the NSI. For instance: the requesting institute or the researcher him/herself have to be on a certified list, or the researcher must belong to an admissible institute. It is assumed that a *bonafide* user is not interested in breaching data confidentiality. Moreover, as technical and/or methodological support is not usually provided it is important that a user is qualified to undertake their research.

Practically, an access request should contain at least:

- the researchers name and position and the institution to which they are affiliated;
- the data to be used in the research;
- a description of the research project;

Best practice

National legislation/organisation is, of course, crucial in determining RDC access procedures. The following summarises best practices currently in use in many countries:

- the research project description should provide:
    - an outline of the scientific research purposes;
    - a preliminary description of the expected results: outputs must be within the scope of the projects;
    - information on how project results will be disseminated, and particularly if they will be made publicly available;
    - motivation for the use of micro data: an explanation of why the research objective can not be realised other than by analysing confidential micro data;

The following issues may be taken into account when developing access procedures, but are strictly dependent on national legislation/organisation and do not reflect minimum standard or best practice, but should be considered by anyone setting up and RDC:

- the application form: This may represent the agreement between RDC and the user or it may be a preliminary step towards a formal agreement. As an application may be returned for revisions and/or clarifications. RDC administrative/technical staff often assist applicants in this preliminary phase.
- entities entitled to access data: Usually, access to RDCs is granted to universities, research institutes or similar entities undertaking research for the public good. In some countries a register of eligible institutions exists). Others use a quite formal procedure to verify eligibility of institutions applying for data access, for instance the Netherlands. Depending on national circumstances, some countries also allow government departments to access confidential statistical data.
- research aims: The aims of a research project should be scrutinised before access is granted. This is to identify projects that are not viable or can be undertaken with less sensitive data, and those that are explicitly intended to embarrass the NSI or damage its ability to make statistics. Any project that disregards the NSI's policies will be rejected (for example some national legislation does not allow RDC-output that is similar to official statistics).

3.4 Access to sampling frames and sensitive variables

RDCs provide access to a range of data, which are likely to have been generated using different sampling frames: for example, business registers may be used for company surveys, while post/zip codes are often used to generate samples for surveys on households/individuals.

By their nature, sampling frames contain identifiable information, such as names, addresses, or even tax reference numbers. A general principle to minimize the risk of disclosure is that such identifiable sampling information, and any other variables that can be used to directly identify individuals should not be made available to researchers. This leaves the characteristics of variables as the only possible source of disclosure, which is why disclosure control procedures are implemented on all output.

By removing sampling frame data, the risk of accidental disclosure is minimised.

Minimum standard

No direct identifiers should be included in data files that are made available to researchers. Names and addresses of individuals and companies, and other identifiers including administrative codes such as health insurance numbers should be excluded.

Where a unique observation reference number is necessary (for example when constructing a panel), the identifier should be replaced by a unique but meaningless reference number.

Depending on national legislation and organisation, some RDCs may choose to provide access to identifying information in special circumstances, for example, the provision of post/zip codes for spatial research.

In all cases, irrelevant of the variables made available the usual disclosure control rules should apply to output.

Best practice

The minimum standard applies.

3.5 Responsibility for quality of outputs

The aim of this section is to leave no doubt as to who is responsible for the content of outputs and the conclusions that are based on these outputs. In an RDC-environment the NSI can not take this responsibility. The aim of an RDC is to allow researchers to undertake analyses for which NSIs do not have the remit and/or resources. For an NSI to take responsibility for the output of these analyses it would imply an involvement in the research project and a responsibility for the conclusions and the quality of output. It should always be made very clear that outputs from an RDC are NOT official statistics.

Minimum standard

For this issue there is no minimum standard. All RDCs should comply with the best practice.

Best practice

Researchers are required to include a disclaimer in all publications, papers, presentations etc, that are (in part) based on research performed at the Research Data Center. The exact wording of the disclaimer can be country-specific but the disclaimer should contain the following elements:

- Data from the NSI has been used in the research, with reference to the exact datasets used
- The presented results are the sole responsibility of the author.
- The presented results do not represent official views of the NSI nor constitute official statistics.

Examples of the exact wording in a few countries are included in annex 1.

3.6 Checking for quality

Almost all RDCs check outputs only for statistical disclosure since quality of output is generally considered the responsibility of the researcher not the RDC. Having said that, many RDCs will give researchers guidance if they believe an output has some quality issues. Also, some RDCs might be tempted to judge an output on its possible negative impact on the NSI itself (for instance outputs that show some quality problems in official statistics).

Minimum standard

An RDC should clearly distinguish between comments relating to confidentiality and to quality. Ideally, comments relating to confidentiality should be delivered formally, while comments on quality are by a more informal channel (for example orally in person or by phone).

Best practice

Only check an output for confidentiality, not for quality or possible negative impact on the NSI itself. If there are concerns over output being used to embarrass an NSI, this should ideally be dealt with during the application phase, when a project can be rejected entirely for this reason (as is outline above). For added security, an NSI could include a clause in their access agreement which requires users to approach the NSI first when they suspect errors or quality issues in the data used.

3.7 Output documentation

Researchers often produce numerous outputs; to understand the data, for reasons of statistical data editing and to specify models. Therefore over time the amount of output which needs to be checked increases. As the majority of the research projects extend over several years and syntaxes may be sent irregularly, checkers need to frequently reacquaint themselves with the details of individual research projects.

This section aims to outline the information that researchers should provide to ensure that the checkers are able to understand and assess complex output quickly and efficiently. The requirements for the documentation of output should be communicated with researchers as part of the access agreement.

Minimum standard

In order to maximise the efficiency of output checking, as a minimum standard all output should include:

- The researcher's name and institute,
- the name (and number where relevant) of the research project,
- the date on which the syntax is submitted,
- a brief description of the purpose of the output,
- the data files used,
- a description of original and self constructed variables,
- in the case of sub samples the selection criteria and the size of the sub sample,
- weighted results and unweighted frequencies.

If the output does not correspond to the minimum standard the checkers are advised to reject the output.

Best practice

For best practice an output should include all the information above, and…

- a full record (log-file) of the analysis,
- a self-explanatory documentation / annotation of the steps of analysis,
- full labelling of all variables and all value labels.

3.8 Checking every output or not?

Creating an efficient output checking process means balancing the time spent with the risk avoided. The extremes are obvious for everyone: no checking leads to large risks and zero personnel costs for checking, while checking everything reduces risk, but generates very high personnel costs for checking. There may be some situations where the time saved by checking only a sample of all outputs outweighs the added risk. Obviously, there is a benefit in this for the researchers as well. It means that they receive most of their outputs immediately, without the time delay that would be necessary for checking them.

Minimum standard

As a minimum standard all outputs are checked. This is to ensure a maximum data security.

Best practice

As a best practice an NSI should define an output checking strategy. As part of this strategy the relative risk and utility of subsampling output should be assessed. Some considerations to take into account when developing this strategy are:

- New researchers should have all their outputs checked. After a set number of non-disclosive outputs have been submitted, a researcher is labelled ‘experienced’.
- It may be possible for experienced researchers to be checked randomly and by subsample. If disclosive output is submitted, the researcher loses their ‘experienced’ status and falls back to the situation where all his output is checked.
- Outputs based on sensitive variables may not be suitable for subsampling.

3.9 Number or size of outputs

The aim of this section is to provide guidelines to prevent the RDC being buried under a pile of outputs. An RDC will typically be used by numerous reseachers at the same time who all want to receive their cleared outputs as quickly as possible. However, output checking capacity is limited, so if researchers submit very many or very large outputs this increases clearing times for RDC users. So an incentive should be built in to ensure that outputs are focused and valuable and that no checking-capacity is wasted. Output checking capacity is a resource that is shared by all users and should therefore be claimed in all fairness.

Minimum standard

The RDC should have (and make known) a policy that allows them to reject any output on the grounds of volume or quantity only, irrespective of its content. This policy should be explained to researchers from the start.

Best practice

A cost barrier (in time or money) is created to prevent many or large outputs. Possible solutions include:

- Allowing each research project only a limited number of free outputs (in the extreme: zero). Researchers would then pay a fixed fee (based on the average time it takes to clear an output) for each additional output.
- Charging an hourly rate for the time spent clearing an output that takes an inordinate amount of time (for instance more that 5 times the average time it takes to clear an output)
- Placing large output at the back of the queue. This means that smaller outputs will be checked first even when they will be submitted at a later point in time. Without such a measure, the 'cooperative researchers' will be punished because they have to wait longer for their output, since the large (and therefore time consuming) output needs to be cleared first.

It should be realised that if a financial barrier is implemented, researchers could start bundling outputs together to form one large output, so they will only pay a fee once. Setting a sensible pricing level (i.e. not too high), could go a long way in preventing this.

3.10 Output clearance times

Most organisations give guidelines but do not make a strict commitment to RDC-users on the time taken to check and return their output. Experience seems to suggest that the majority of output can be cleared in five working days. However, this depends both on the type and the size of the results submitted, as well as RDC staff resources. Defining a fixed timeframe in which researchers may expect to receive cleared output may be seen as a pressure on checkers, but the aim has to be to avoid uncertainty among users over how long they can expect to wait for output.

Minimum standard

Provide users with an indication of the average time needed to check the output with reference to the type and the volume of the submitted output, without making a commitment when this specific output will be checked.

Best practice

Response times should be monitored, and exceptions should be documented.

A commitment should be made on the maximum time a user can expect to wait for a response, either in the form of a released output or a request for further information.

It is worth noting that response time can be influenced by the way the RDC is funded. If a fee is charged for output checking, it limits the amount of output submitted and makes the users more aware of what they really need as output but, on the other side, it places an obligation on the NSI to reply in a specified timeframe.

3.11 Number of checkers

The aim of this guideline is to protect the RDC against mistakes and to ensure confidentiality of outputs. The guideline also ensures that the output checks are consistent across all checkers.

Minimum standard

As a minimum standard the output should be checked by one employee. The RDC has to ensure that the output corresponds to the legal rules for confidential data.

Best practice

The best practice is the ‘four eyes principle’. Two options exist for implementing the four eyes principle. First, the two checkers are both staff members of the RDC and second the first check is accomplished by one RDC employee (expert in statistical analyses) and the second check is done by the subject-matter department (expert in

data). This last procedure lowers the risk of disclosive output being released, because the two checkers each bring their own expertise and in that way complement each other.

3.12 Skill of checkers

Researchers often use complex statistical methods for their analyses, which generate a wide range of statistical and econometric outputs. Because output checking is time consuming and costly it is vital for efficiency that checkers fulfil a minimum standard of skill. The aim of this guideline is to ensure that checking is accurate, consistent and that checkers do not waste time understanding data and statistical methods.

Minimum standard

The RDC has to ensure that the employees fulfil the following requirements

- the technical expertise to understand most of the output,
- a working knowledge of the data held in the RDC,
- periodic training in new statistical methods combined with regular reviews

Best practice

In addition to the requirements above, for best practice,

- at least one checker should have active research experience,
- output checking should form the main part, but not all of checker's job. This enables RDCs to employ qualified staff who might otherwise be uninterested in a job that consisted only of checking other people's work.
- To facilitate the training and review process mentioned above, a copy of all outputs and discussions relating to those outputs should be archived.

4. Researcher training

Training for RDC-users (use of the facility, legal aspects, disclosure control, etc.) is widely considered a good practice. Usually, a face-to-face contact is used to train the users, and this helps in building trust and increase security

Minimum standard

For the minimum standard face to face training is not necessary, however documentation should be provided covering:

- the researcher's legal responsibilities
- instructions on how to use the facility
- disclosure control principles to ensure researcher understand the issues

Best practice

For best practice a standard face-to-face training module should be developed covering the issues outlined above. It may be designed for group or individual presentation, but the criteria for such a module would be that:

- it should give researchers basic tools to use the facility without assuming that they read the documentation
- it can be delivered by more than one person without significant loss of consistency

Training topics

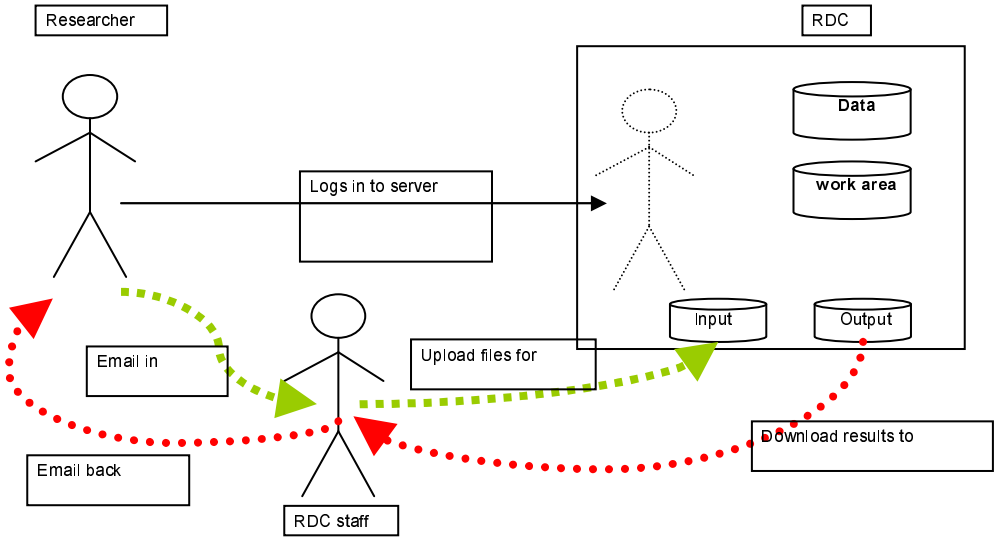
Since the organisational and procedural aspects of running an RDC are country specific, these guidelines outline the topics that should be included in best practice researcher training. The exact format of a training course can be tailored to a country's needs.

Part 1: Process

- Legal and ethical background
    - Laws and regulations governing RDC
    - Code of ethics
- The Research Data Centre (RDC): role and purpose
    - NSI has a duty to support research
    - concerns about confidentiality risks

- RDC security model: principles
    - valid statistical purpose — safe projects
    - trusted researchers — + safe people
    - anonymisation of data — + safe data
    - technical controls around data — + safe settings
    - disclosure control of results — + safe outputs

      ⇒ safe use
- Who can access the RDC
- RDC use:
    - diagram of interactions (user/data/staff/output), for example



- What users cannot do:
    - attempting to remove data (including writing down data from the screen)
    - attempting to identify individuals, households, or firms
    - using data which they are not allowed to
    - using data for anything other than the proposed project
- Penalties and disciplinary actions:
    - In violation of laws
    - In violation of RDC rules / agreement
- RDC House rules
    - booking
    - charges and payment

    - user's behaviour (keep workspace clean, do not misuse resources,…)
    - output description
    - output releasing
- Applying for access:
    - Research project description (including preliminary output description)
    - Details of people applying for access
    - Legal agreement with the institution and the researcher
    - How to submit applications / time frame
- Role of the RDC team

Part 2: Disclosure control

- General principle of disclosure control
    - finding out confidential information
    - associating that information with its source
- Key concepts
    - Two approaches: rules of thumb and principles based
    - Safe/unsafe classes of output
    - Primary disclosure
    - Secondary disclosure
- Rules
    - Rules of thumb for each class of output (with examples)
    - Extra attention on tabular output
        - Primary/secondary disclosure
    - Recommendations and hints for safer output
- Practicalities of disclosure control at RDC
    - Expectations on clearing times
    - Output description
    - Number and size of outputs

Annex 1. Examples of disclaimers

Researchers are required to include a disclaimer in all publications, papers, presentations etc that are (partly) based on research performed at the Research Data Center. The exact wording of the disclaimer can be country-specific but the disclaimer should contain the following elements:

- Data from the NSI has been used in the research
- The presented results are the sole responsibility of the author.
- The presented results do not represent official views of the NSI nor constitute official statistics.

Examples of the exact wording in a few countries are included below.

UK

This work contains statistical data from ONS which is Crown copyright and reproduced with the permission of the controller of HMSO and Queen's Printer for Scotland. The use of the ONS statistical data in this work does not imply the endorsement of the ONS in relation to the interpretation or analysis of the statistical data. This work uses research datasets which may not exactly reproduce National Statistics aggregates."

Italy

The data used in the present work stem from the Italian National Statistical Institute (Istat) – survey xxxx. Data has been processed at Istat Research Data Centre (Laboratorio ADELE).

Results shown are and remain entirely responsibility of the author; they neither represent Istat views nor constitute official statistics.

Annex 2. Output description

Each output should be described by the researcher. This description should contain at least the following items:

- Researcher's name
- The date on which the output is submitted
- The research project that the output belongs to
- Brief description of the purpose of the output
- The datafiles used

The following items could be added if need be:

- Any relation to earlier outputs (for instance, if it is a small adaptation of a previous output)
- Name and location of the outputfile
- Email address to sent the output to after clearing it
- The company that the researcher works for

Demands on the output itself

- Full labelling of all variables (including self constructed ones)
- A description of self constructed variables (specifying the analytical transformation or recoding applied)
- Use of subsamples/subpopulation: specify the selection criteria and size of the subsample/subpopulation
- Use of weights: show weighted and unweighted results
- For magnitude tables, report the corresponding frequency table
- For graphs, report the underlying data (or better yet, just the underlying data without the graph)
- For indices, report the formula and each single factor composing the index
- No hidden items (for instance hidden columns in excel, hidden data behind an excel graph, hidden tables in Access etc)
- Analytical results separated from descriptive tables

Annex 3. Glossary

(most definitions were taken from http://neon.vb.cbs.nl/casc)

NSI	National Statistical Institute
Microdata	A microdataset consists of a set of records containing information on individual respondents or on economic entities
Research Data Center	A common term for the part of an organisation that is responsible for providing access to their microdata for research purposes.
On Site	A facility that has been established on the premises of an NSI. It is a place where external researchers can be permitted access to potentially disclosive data under contractual agreements which cover the maintenance of confidentiality, and which place strict controls on the uses to which the data can be put. The on-site facility can be seen as a ' safe setting' in which confidential data can be analysed. The on-site facility itself would consist of a secure hermetic working and data storage environment in which the confidentiality of the data for research can be ensured. Both the physical and the IT aspects of security would be considered here. The on-site facility also includes administrative and support facilities to external users, and ensures that the agreed conditions for access to the data were complied with.
Remote Access	A facility very similar to On Site. The only difference is in the access itself. For On Site, users have to come to the premises of the NSI, while for Remote Access, researchers can access the facility over a secure internet connection. Sometimes tokens or biometric devices are used for added security when logging on.
Remote Execution	Submitting scripts for execution on microdata stored within an institute's protected network. The submitter writes the scripts using the metadata of the original microdata file or with the use of a dummy dataset with the same structure as the original microdata file. Results of the script are checked against disclosure.
Output checking	The process where research results (tables, models, estimations etc) that researchers have created and want to take out of the controlled environment of the RDC are checked for possible disclosure. Only when found non-disclosive, they are then sent to the researchers.

Disclosure	Disclosure relates to the inappropriate attribution of information to a data subject, whether an individual or an organisation. Disclosure has two components: identification and attribution. Identification: Identification is the association of a particular record within a set of data with a particular population unit. Attribution: Attribution is the association or disassociation of a particular attribute with a particular population unit.
Statistical disclosure control (SDC)	Statistical Disclosure Control techniques can be defined as the set of methods to reduce the risk of disclosing information on individuals, businesses or other organisations. Such methods are only related to the dissemination step and are usually based on restricting the amount of or modifying the data released.
Threshold rule	Usually, with the threshold rule, a cell in a frequency or magnitude table is defined to be sensitive if the number of contributors to the cell is less than some specified number. When thresholds are not respected, an agency may restructure tables and combine categories or use cell suppression, rounding or the confidentiality edit, or provide other additional protection in order to satisfy the rule. The threshold rule can be set on weighted or unweighted cell counts.
Group disclosure	Group disclosure is the situation where some variables in a table (usually spanning variables) define a group of units and other variables in the table divulge information that is valid for each member of the group. Even though no individual unit can be recognized, confidentiality is breached because the information is valid for each member of the group and the group as such is recognizable