Statistical Disclosure Control *for dummies*

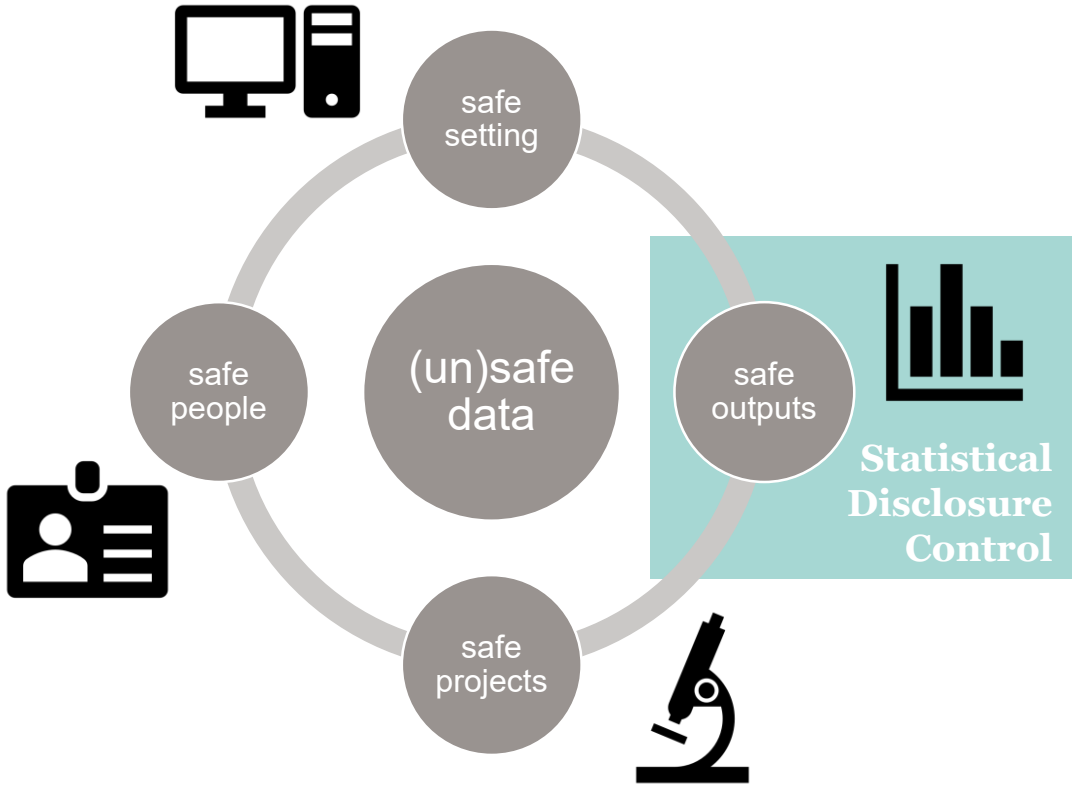
Carlotta Greci

1st June 2018 IASSIST, Montreal (CA)

The Health Foundation

Contents

5 Safes model



Desai, T.; Ritchie, F.; Welpton, R. (2016). "Five Safes: designing data access for research". Bristol Business School Working Papers in Economics

Statistical Disclosure Control

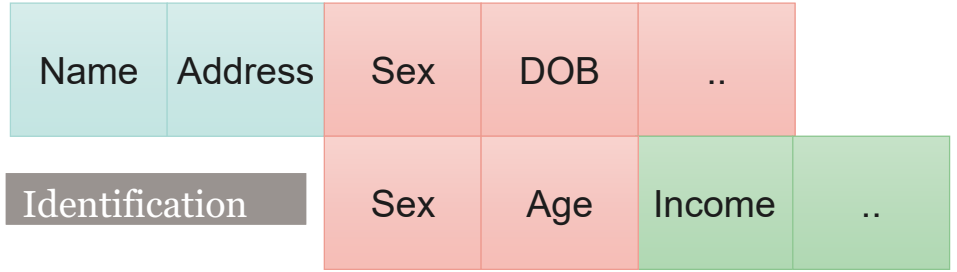
SDC as a tool to mitigate risk

i.e. practice to *reduce* the risk of a disclosure

Can apply to any statistical output

Focus on quantitative methods

- Identification
- Attribution
- *Secondary disclosure*



Secondary disclosure

Income levels for TH				by gender
	0-35k	[illegible]	>65k	Total
Female	49 (-1)	100	0	150
Male	5	50	100	155
Total	55	150	100	305

Attribution

What is the problem?

Not an exact rule exercise: it is a risk assessment!

- Existing guidance developed for tabular outputs
- “Old” material - novel methodologies?

- Lack of consistency across disciplines
- Need for something practical to support practitioners
- National accreditation from UKSA & ONS
- Evolving data privacy legal scenario

Who we are

We are part of the working group for **Safe Data Access Professionals** in the UK

Cancer Research UK
Richard Welpton

UK Data Service
Christine Woods
James Scott

The Health Foundation
Arne Wolters
Carlotta Greci

The project

Create a practical **Handbook** for practitioners

- not prescriptive but informative
- specific disclosure risk for each type of output
- alternative mitigating options
- aim towards the release of the output!

Tips for organisations on managing SDC process

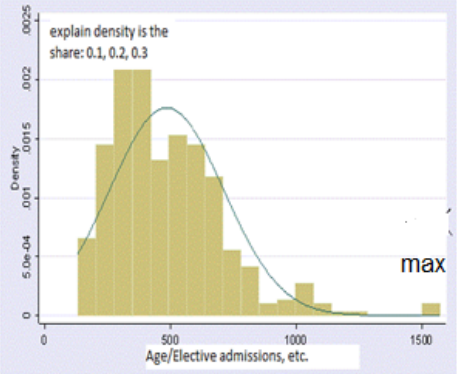
Guidance for analysts on producing good outputs

Assessing disclosure (1)

Histograms

Histograms and density plots

Figure 4 – Example of age distribution/



Minimum info

table of underlying frequencies (e.g. no of data subjects associated to each bar)
Labels for axis & variables + title
Specs of data subjects and total counts

A histogram plot displays the frequency distribution of a variable, where the width of the bars can either represent class intervals or a single value and the eight the frequency density. A density plot shows the distribution of the data over a continuous or discrete variable and it is often used to display the shape of the distribution of a specific variable.

In the example, the chart shows the frequency distribution of age where each histogram shows the percentage of data subjects falling within each age class (e.g. 20% of workers are between 20-30 years old).

Assessing disclosure (2)

Histograms

Where the disclosure lies

SDC considerations

- Graphs should abide by the same SDC rules for tabular outputs. Ideally, the easiest way to check graphs is to apply SDC to the underlying table(s) generating the chart. For histograms and density plots, common SDC issues arise from low cell counts (below N data subjects) and min/max values.
- As histograms and density plots are used to show the distribution of a value, low counts are often an issue, especially on the tails. Min and max are often masked in the scale of the X-axis as the start and ending points. Many statistical software (e.g. Stata) use by default the max value as the ending value for the X-axis.
- Graphs should be released as fixed images (e.g. JPEG), as some statistical software (e.g. Excel) can store data behind a graph. If a user needs to recreate the graph in a particular layout or format they should use underlying frequencies once they are checked for SDC.

Rule of thumb

Same rules apply as for frequency tables and min and max values.

Assessing

Histograms

Where the disclosure lies

How to release the output

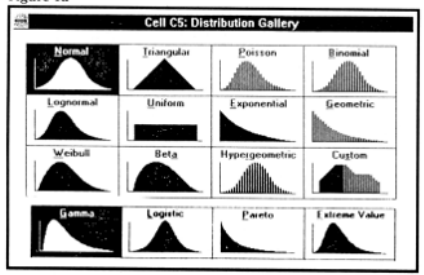
Reducing disclosure risk

It is important that analysts specify the purpose of the graph, as the mitigating options may vary depending on the meaning behind the output. If both chart and frequency table are released, the same mitigating actions should apply to both.

1. If the graph is intended to show that the distribution has a long tail (i.e. the presence of many outliers), it is likely that the low count observations are concentrated in the upper or lower part of the distribution. The preferable option would be to cap all these values in one class. This approach can also be used to mask the maximum or minimum values (see example below). This is not necessary in cases where the minimum or maximum are a structural value or they are defined by the analyst. In this example, the minimum is a structural zero as there cannot be a lower value for age and therefore no need to conceal it. The same would apply if the data subjects were selected within an age band (i.e. only individuals less than 65 years old), where the upper limit would coincide with the max value of 65.

Figure 5

Figure 1a



2. If the aim is to show the shape of a probability distribution, it is possible to keep the full plot by omitting all values on the X-axis (see figure 3). With this solution, it is possible to relax the rules of thumb for low counts and min and max, as it would not be possible to associate any value to a specific bar or point in the graph.

3. If the low counts are distributed outside the tails, it may be a good solution to band the

Types of outputs

- Descriptive stats
- Box plot
- Concentration ratios
- Exclusion criteria
- Factor analysis
- Histograms & density plots
- Gini coeff
- Margin plots
- Percentiles
- Regressions & test stats
- Residuals
- Risk stratification
- Scatter plots
- Symmetry plots
- Stand differences
- Spatial analysis
- Survival analysis
- Time series

For organisations..

Design SDC process to fit the organisation's needs & risk appetite

Some tips:

- encourage good outputs
- independence of checkers
- 4 eyes principle
- workload & pressure
- accountability & auditing

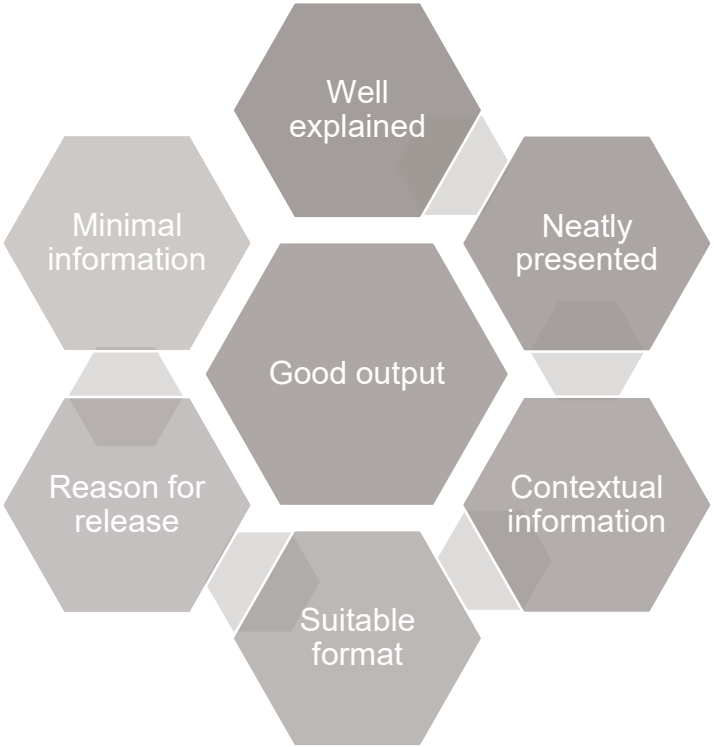
The key is.. good outputs!

What does make an output good?

- understanding of SDC principles
- be aware: cannot eliminate full risk of disclosure

Engaging with analysts

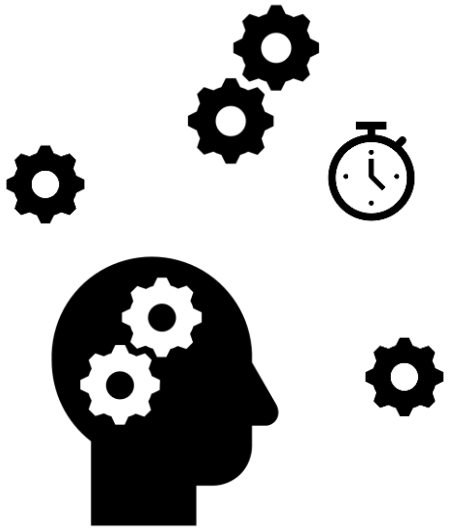
- SLA
- Consistency
- Training



What is next

- Developing a training dataset & material
- Review with SDC practitioners
- External peer review
- Informing national consultation (UKSA & ONS)
- Publication (expected August 2018)

- Sharing it with our IASSIST colleagues!



Thank you

Arne, Carlotta, Christine, James & Richard

Carlotta.Greci@health.org.uk

The Health Foundation