GSS/GSR Disclosure Control Guidance for Tables Produced From Administrative Sources – Case Studies

Case Study 1

Table 1 shows female benefit claimants at a low geography along with age characteristics of the individual. An individual could be identified by spontaneous recognition or through malicious intent Attribute disclosure has occurred if someone recognises someone they may know e.g. their neighbour and discovers from the table that they are claiming this benefit.

Disclosure may arise if there is a count of 1 in a marginal total (row or column) as in the simple example shown in table 1, where for example claimants for Sickness, Child and Housing benefit are broken down by age bands. Anyone who knows that a particular individual between 16 and 24 years receives a benefit would learn that it was Sickness benefit. Attribute disclosure could occur from a count of 2 in a marginal total where one of the units may identify the other and thereby disclose further information.

Table 1: Benefit claimants, by type and age (Females only)

Age	16-24	25-49	50-59	>59	Total
Sickness benefit	1	0	7	1	9
Child benefit	0	0	18	19	37
Housing benefit	0	12	5	0	17
Total	1	12	30	20	63

Attribute disclosure can also occur from cells with larger values, where they appear in a row or column dominated by zeros. A zero in population data allows one to say that no-one in the population has that attribute. This can be seen in Table 1, which reveals that as no 25-49 year olds are claiming sickness or child benefit all 25-49 year old females in the table are claiming housing benefit. The risk from many zeros within tables may not be significant, but, in some cases, they may need to be protected. This can depend on the distribution of the zeros and whether they dominate a row or column.

Disclosure risks may increase where groups of units who appear in the same table know enough about each other to identify each other and potentially discover something new. This can occur where units share characteristics or are grouped in some way, e.g. individuals from the same household.

In order to protect against general attribute disclosure, at a minimum, care should be taken where rows or columns are dominated by zeros and in particular where a marginal total is a 1 or 2.

A possible solution is collapsing the age categories as in Table 1a.

Table 1a: Benefit claimants, by type and age (Females only)

Age	16-59	>59	Total
Sickness benefit	8	1	9
Child benefit	18	19	37
Housing benefit	17	0	17
Total	43	20	63

Much detail has been lost here but still low counts (0 and 1) remain. Recoding age into 2 groups does not solve any disclosure issues. In many cases combining categories can be a successful approach although any loss in data utility needs to be kept in mind.

Another solution is suppression as shown in Table 1b. Cells of value 0) and 1 are suppressed while other cells are selected for secondary suppression.

Table 1b: Benefit claimants, by type and age (Females only)

Age	**16-24**	**25-49**	**50-59**	**>59**	**Total**
Sickness benefit	c	c	7	c	9
Child benefit	c	c	18	c	37
Housing benefit	c	c	5	c	17
Total	1	12	30	20	63

Due to the nature of this table many cells require suppression, either primary or secondary. The table is safe but little information of use remains. It is of slightly greater utility if the marginal totals for the two lower age groups are published although this would require further thought as the risk of disclosure would be increased.
In the example in Table 1b all cell counts of zero have been suppressed. However some consideration is needed on the issue of whether always to suppress counts of zero. Clearly where sufficient numbers of zeros can expose other counts and potential disclosures, zeros need to be suppressed. One would have to consider whether by including any zeros in a table one is effectively creating an attribute disclosure. In doing this, the impact of any such 'disclosure' must be assessed - how likely would it be to allow someone to find out something about an identified individual with any certainty, and would the information discovered be sensitive? In addition would information loss be increased by including zeros? The relationship between utility and confidentiality would need to be considered.

A further solution here is to release the data at a higher level of geography. The table could possibly be released without any disclosure control being applied although some data utility is lost by producing the table at this higher level.

Changing the age group categories is also an option although in many cases this would not be practical as standard categories would be expected.

Case Study 2

An intruder with a special interest in conception statistics discovers from Table 2 that a small number of twins have been born to mothers in specified age groups in England and Wales. The small number in the cell does not lead to direct identification but it may prompt an intruder to use knowledge of the age of an acquaintance giving birth to find out that she had given birth to a still born twin in addition to a living child.

Table 2: Number of twins by age group (adapted from Birth Statistics FM1 Table 6.4)
LM - Liveborn male SM - Stillborn male LF - Liveborn female SF - Stillborn female

England and Wales		Age of Mother at birth of child							
		Under 20	20-24	25-29	30-34	35-39	40-44	45 and over	Total
Outcome of birth	2 LM	93	356	678	982	345	98	921	3,473
	1 LM and 1 LF	32	234	589	1,001	612	121	34	2,623
	2 LF	78	345	865	943	532	24	103	2,890
	1 LM and 1 SM	3	10	12	11	15	3	1	55
	1 LM and 1 SF	1	3	3	5	3	0	1	16
	1 LF and 1 SM	3	1	0	8	3	2	0	17
	1 LF and 1 SF	0	2	4	3	4	4	2	19
	2 SM	0	3	2	3	0	0	0	8
	1 SM and 1 SF	1	1	0	0	1	1	2	6
	2 SF	0	1	6	2	3	0	1	13
	Total	211	956	2,159	2,958	1,518	253	1,065	9,120

There are many cells in this table with low counts relating to sensitive information. A motivated intruder having been told informally that an acquaintance aged 45 and over was pregnant with twins could later investigate further to discover the outcome (e.g. if aware that mother later only had one child, a female, the stillborn child must have been female too). There is no register of stillbirths so the information is not in the public domain.

These low counts, including zeros, require protection. Table 2a shows a combination of combining age categories and suppression.

Table 2a: Number of twins by age group (adapted from Birth Statistics FM1 Table 6.4)
LM - Liveborn male SM - Stillborn male LF - Liveborn female SF - Stillborn female

England and Wales		Age of Mother at birth of child			
		Under 25	25-34	35 and over	Total
Outcome of birth	2 LM	449	1,680	1,364	3,473
	1 LM and 1 LF	266	1,590	767	2,623
	2 LF	423	1,808	659	2,890
	1 LM and 1 SM	13	23	19	55
	1 LM and 1 SF	4	8	4	16
	1 LF and 1 SM	4	8	5	17
	1 LF and 1 SF	2	7	10	19
	2 SM	3	5	0	8
	1 SM and 1 SF	c	c	4	6
	2 SF	c	c	4	13
	Total	1,167	5,117	2,836	9,120

There is still a good deal of practical utility in Table 2a although some of the finer detail has been lost. It would need to be assessed as to whether this table would be sufficient for research purposes.

If detail on still births was not required the data could be released as in Table 2b although suppression may be required to protect the cells with a value of 1. Given the table is at England and Wales level, the only risk is likely to be self-identification rather than any attribute disclosure, but the high sensitivity of the table may persuade a data provider to apply protection. Risk categories of particular variables (including health data) are discussed in A3.4. In summary a table such as Table 2b is likely to be considered sensitive as any personal information obtained from this table could have an extreme effect on the individual concerned. Therefore it is less likely to be released as it stands.

Table 2b: Number of twins by age group (adapted from Birth Statistics FM1 Table 6.4)
LM - Liveborn male SM - Stillborn male LF - Liveborn female SF - Stillborn female

England and Wales		**Age of Mother at birth of child**							
		Under 20	**20-24**	**25-29**	**30-34**	**35-39**	**40-44**	**45 and over**	**Total**
Outcome of birth	**2 LM**	93	356	678	982	345	98	921	3,473
	1 LM and 1 LF	32	234	589	1,001	612	121	34	2,623
	2 LF	78	345	865	943	532	24	103	2,890
	1 Living and 1 Stillborn	7	16	19	27	25	9	4	55
	2 Stillborn	1	5	8	5	4	1	3	8
	Total	211	956	2,159	2,958	1,518	253	1,065	9,120

This is an alternative recoding approach to that in Table 2a. The method of recoding would reflect the requirements of the users. Due to the large number of small cells rounding was not considered for Table 2 as it would result in a large proportion of cells being rounded to zero.

Case Study 3

A table of Live Births outside marriage/civil partnership in geographical area below PCO shows a count of 1 when the age groups of both parents are shown. The couple can see that they are the only combinations of these age groups to have had a child in this year. This knowledge could make the couple feel especially vulnerable. An example is seen in Table 3.

Table 3: Number of births by age of parents

		Age of Mother at birth of child							
		< 20	**20-24**	**25-29**	**30-34**	**35-39**	**40-44**	**> 44**	**Total**
Age of Father at Birth of child	**< 20**	634	145	65	18	9	2	0	873
	20-24	1,326	1,452	764	346	89	23	1	4,001
	25-29	314	1,264	1,045	890	649	257	26	4,445
	30-34	62	746	784	651	451	365	38	3,097
	35-39	21	259	476	328	214	169	17	1,484
	40-44	13	214	331	267	135	83	15	1,058
	45-49	6	46	54	41	36	23	8	214
	50-54	1	11	21	32	17	11	6	99
	55-59	1	3	6	12	8	6	4	40
	60-64	0	2	3	4	3	4	1	17
	> 64	0	0	1	1	0	1	0	3
	Total	2,378	4,142	3,550	2,590	1,611	944	116	15,331

An example of disclosure for this table could occur if an attacker attempts to identify the parents of a child where there is a large gap between the ages of the father and mother, such as the count of 1 for a father aged over 65 and the mother aged 25-29.

As with the previous case studies, the table can be protected by collapsing categories as can be seen in Table 3a.

Table 3a: Number of births by age of parents

		Age of Mother at birth of child						
		< 20	**20-24**	**25-29**	**30-34**	**35-39**	**> 39**	**Total**
Age of Father at Birth of child	**< 20**	634	145	65	18	9	2	873
	20-24	1,326	1,452	764	346	89	24	4,001
	25-29	314	1,264	1,045	890	649	283	4,445
	30-34	62	746	784	651	451	403	3,097
	35-39	21	259	476	328	214	186	1,484
	40-44	13	214	331	267	135	98	1,058
	45-49	6	46	54	41	36	31	214
	> 49	2	16	31	49	28	33	159
	Total	2,378	4,142	3,550	2,590	1,611	1,060	15,331

If the cells of size 2 are considered disclosive then suppression could be applied. Another option would be to collapse the table further and create categories 45 and over (for the father) and 35 and over (for the mother).

Rounding to base 3 could be an option here although there would be considerable loss of data utility as many cells have low counts. This can be seen in Table 3b.Simple rounding is applied where each cell is rounded to the nearest multiple of 3. This often results in the table being non additive. Consideration would need to be given in this example as whether to replace the dashes with zeros. Keeping this distinction will make it easier for an attacker to unpick the table as it would be obvious to a researcher that cells with counts of zero had been rounded down from a larger number. Table 3b is shown with both dashes and zeros displayed to allow direct comparison with the unrounded table.

Note that the latest coding convention as used by Eurostat requests that '0' can only be used when the figure is a true zero, not when a figure has been rounded to zero. In these cases '0~' should be used. As discussed in the previous paragraph, this will affect the confidentiality of the table.
An exception can be made for structural zeros (cells for which an event cannot occur, such as 0-5 year olds being a mother.

Table 3b: Number of births by age of parents

		Age of Mother at birth of child							
		< 20	**20-24**	**25-29**	**30-34**	**35-39**	**40-44**	**> 44**	**Total**
Age of Father at Birth of child	**< 20**	633	144	66	18	9	3	-	873
	20-24	1,327	1,452	765	345	90	24	0	4,002
	25-29	315	1,263	1,044	891	648	258	27	4,446
	30-34	63	747	783	651	450	366	39	3,096
	35-39	21	258	477	327	213	168	18	1,485
	40-44	12	213	330	267	135	81	15	1,059
	45-49	6	45	54	42	36	24	9	213
	50-54	0	12	21	33	18	12	6	99
	55-59	0	3	6	12	9	6	3	39
	60-64	-	3	3	3	3	3	0	18
	> 64	-	-	3	0	-	0	-	3
	Total	2,379	4,143	3,549	2,589	1,611	945	117	15,330