Assessment of Data Protection risk in PIs

PITG 13/02

Issue

1. An assessment of whether current disclosure control techniques used in PIs are adequate and whether moving to a percentage suppression threshold of 22.5 in the 2013 PIs would provide greater security.

Conclusion and recommendations

2. The current disclosure control techniques of rounding of counts and suppression limits for percentages coupled with the design of the tabulations result in a very low risk that personal information can be deduced from the existing PIs. A move to a percentage threshold of 22.5 would not have a significant impact on the level of risk, though this may be advantageous for other reasons.

3. The discussion below raises some questions for PITG members to consider, which could provide a still greater level of confidence. These relate to the possible increasing of thresholds for certain categories of characteristic that represent small minorities of the student cohort, and the reduction in precision of some percentages. PITG members must balance the marginal increase in disclosure control with the consequent loss of detail in the indicators, either real or perceived.

Discussion

4. Any publication of data relating to individuals creates a risk to the privacy of those individuals. Depending on the type of data, and the size of the populations concerned, this risk might be very low, for example where all counts are large numbers. However where data deals with small numbers, and where the subject of the data is in any way sensitive, the privacy risk may be more significant.

Data suppression in HE Performance Indicators (PIs)

5. The intention of data suppression in the PIs is to reduce the risk of identifying individuals from the published data. The PIs apply two types of disclosure control. Counts are rounded to the nearest multiple of 5. Percentages, including the indicators themselves, are suppressed if based on a population of 19 or fewer. This is a lower threshold than normally applied to HESA outputs where percentages are suppressed if based on a population of 52 or fewer. The lower threshold allows more indicators to be published.

6. The increased risk of disclosure from the use of a lower threshold is difficult to quantify. The HESA standard threshold is designed to be applied to any novel extract with the same confidence of minimal risk. The possibly higher risk of a lower threshold for PIs is mitigated by their predictability – the same indicators are calculated each year and novel cross-tabulations are not possible with the published data. Whether the population threshold is set at 20 or 52, or any value in between, it is possible in theory to deduce information that describes one individual. However this does not mean that one could necessarily identify who that individual is and learn new information about them.

Finding a single individual - example

7. Table T1a displays the number of young full-time first degree entrants and of those the number from state schools (each rounded to nearest 5), and displays the indicator percentage to one decimal place as follows:

	no. with known data	no. from state schools	indicator
University of Poppleton	20	20	95.0%
Poppleton College	20	20	95.2%
Poppleton Metropolitan	55	50	98.1%

8. The rounded counts could represent any of the following possible combinations of un-rounded raw data, of which only one combination would result in each of the three indicator percentages:

no. with known data	no. from state schools	indicator	no. with known data	no. from state schools	indicator
20	18	90.0%	54	51	94.4%
20	19	95.0%	54	52	96.3%
20	20	100.0%	55	48	87.3%
21	18	85.7%	55	49	89.1%
21	19	90.5%	55	50	90.9%
21	20	95.2%	55	51	92.7%
22	18	81.8%	55	52	94.5%
22	19	86.4%	56	48	85.7%
22	20	90.9%	56	49	87.5%
53	48	90.6%	56	50	89.3%
53	49	92.5%	56	51	91.1%
53	50	94.3%	56	52	92.9%
53	51	96.2%	57	48	84.2%
53	52	98.1%	57	49	86.0%
54	48	88.9%	57	50	87.7%
54	49	90.7%	57	51	89.5%
54	50	92.6%	57	52	91.2%

In each case we can deduce that each institution had only one young full-time first degree entrant from an independent school. The same method can apply to any of the Widening Participation measures, the T3 Non-continuation rates and the E1 Employment of leavers tables.

Can one identify an individual?

9. In the above example one cannot learn new information about the single entrant from an independent school. One can only identify who the single entrant is if one already knows that she attended an independent school, and hence the data has revealed nothing new. However, if one did already know that Jane Bloggs had attended Poppleton Ladies College before entering Poppleton Metropolitan University, then one can deduce that every other entrant had attended a state school (assuming there are no unknowns – which could also be deduced by the same method). The suppression thresholds do mean that no two tables in the PIs can be compared to gain additional information about an individual. For example one could deduce, from T1a, that a single entrant at Poppleton College came from a Low Participation Neighbourhood, but one could not then learn about their continuation status from the following year's T3b (non-continuation of LPN entrants) because the percentage calculation would be suppressed.

What is the risk? Is it acceptable or is further mitigation necessary?

10. The examples show that cases can arise in the PIs where one can deduce that a single individual at an institution meets a particular description. However these cases only arise if:
    - There is in fact only one individual,
    - The combination of rounded counts and the percentage calculation reveal only one possible combination of un-rounded raw figures,
    - Those figures differ by one,
    - It can also be deduced that there are 0 unknowns.

11. This combination of factors is not likely to occur often, and if it does one still cannot learn new information about an individual. It is also worth considering whether anyone would have the desire or motivation to make this deduction if such a combination did arise. The likelihood of this combination of factors occurring is marginally reduced by using a higher threshold population for suppression because smaller populations are more likely to have only one individual that meets a particular description. The case of Poppleton Metropolitan above, with a population >52, shows that the combination of factors could still occur.

12. Raising the threshold to 22.5 would not significantly reduce the likelihood of the risk factors occurring, but could potentially reduce the number of indicators that can be published. The group needs to balance the risk described above with the usefulness of the data. Changing the precision with which percentages are displayed might reduce the risk slightly, but the Poppleton Metropolitan again shows a case where the identifiable combination arises even after rounding to 0dp (only one value rounds to 98%). It is reasonable anyway to query whether differences of one in a thousand are relevant when dealing with populations of less than a hundred.

13. This final example shows the combination of factors arising with a population greater than 22.5. With indicators shown to 0dp in this example one cannot deduce the exact un-rounded raw data, but one can still deduce a single independent school entrant:

	no. with known data	no. from state schools	indicator
Poppleton Polytechnic	25	25	96%

no. with known data	no. from state schools	indicator to 1 dp	indicator to 0 dp
23	23	100.0%	100%
24	23	95.8%	96%
24	24	100.0%	100%
25	23	92.0%	92%
25	24	96.0%	96%
25	25	100.0%	100%
26	23	88.5%	88%
26	24	92.3%	92%

no. with known data	no. from state schools	indicator to 1 dp	indicator to 0 dp
26	25	96.2%	96%
26	26	100.0%	100%
27	23	85.2%	85%
27	24	88.9%	89%
27	25	92.6%	93%
27	26	96.3%	96%
27	27	100.0%	100%

14. The threshold of suppression could be determined by assessing the actual chance of there being just one member of a group. Very roughly the overall UK proportions of students in each minority group are as follows:

Independent schools	12%
NS-SEC 4-7	30%
LPN (POLAR 2)	10%
DSA	5%
Non-continuation	7%
Unemployed	10%

15. Therefore it is fairly likely that a population of 20 will have exactly one DSA recipient or one non-continuer, but probably more than one of each other category. There may therefore be an argument for raising the percentage threshold for DSA and non-continuation PIs to 40 or even 50, but not for the others. However, as stated above the fact that one individual may be deduced from the counts does not in itself present a problem if no new information can be deduced about that individual.

Further information

For further information contact Simon Kemp (tel: 01242 211122; e-mail: simon.kemp@hesa.ac.uk).