Paper PAP_PP11

De-identification of Participant Data for Public Disclosure

Nicola Perry, GSK, London, UK

ABSTRACT

k-anonymity is a property possessed by certain anonymized data. The concept of k-anonymity was first introduced by Latanya Sweeney and Pierangela Samarati in a paper published in 1998. A release of data is said to have the k-anonymity property if the information for each person contained in the release cannot be distinguished from at least k - 1 individuals whose information also appear in the release. According to Health Canada k should be 11. A public sharing threshold of 0.09 indicates that a person's information cannot be shared unless it cannot be distinguished from at least 10 other individuals within the same release. This paper aims to explain how GSK have applied this threshold in data displays that have a number of participants with direct identifiers (age/race/ethnicity) for studies in scope of public disclosure to minimize the possibility of re-identification of participants.

INTRODUCTION

It is essential that the privacy of participant data in clinical trials is maintained. Although data displays produced for public disclosure do not contain personally identifying information, e.g., participant's name, they may include other distinctive data such as age and sex. If this data is combined and linked to publicly available data, there is a possibility participants could be re-identified. In order to minimise the possibility of re-identification of participants, additional tables are required for the purpose of data disclosure. As per k-anonymity concept, the key number in any individual table cell is 11. Counts below this must be considered a risk to re-identification.

Data disclosure summaries posted to clinical trial registers, e.g., ClinticalTrials.gov, contains tabulations of participant demographic characteristics. This paper aims to explain GSK's approach to the de-identification of the potential identifiers within this tabulation.

APPROACH TO DE-IDENTIFICATION OF DEMOGRAPHIC CHARACTERISTICS

POTENTIAL IDENTIFIERS

The following variables are always considered potential identifiers:

- Age (continuous and ordered categorical).
- Sex (categorical).
- Race (categorical).
- Ethnicity (categorical).
- Region / Country (categorical).
- Site (categorical).

Any variable or combinations of variables which leads to a non-zero count (n) of <11 participants in a clinical study having this mix of characteristics must be considered as a risk to re-identification.

DE-IDENTIFICATION OF POTENTIAL IDENTIFIERS

For all potentially identifying categorical variables, all non-zero cells within a table must have n≥11. In addition, for continuous variables, the standard deviation must be >0.

If these criteria are not met, the steps in Table 1 are carried out.

TABLE 1

Type	Examples of Variables	Steps
Non-ordered categorical variables	Sex. Race. Ethnicity. Region / Country. Site.	**Step 1**: Combine all categories for which at least 1 cell has 0<n<11, and label this new, combined category as 'De-identified'. **Step 2**: If all cells within the de-identified category have n≥11, then stop. Else combine the de-identified category with the category with the lowest overall n (n>0). This new combined category is labelled 'De-identified'. In the case of 2 or more categories having equal lowest total n, combine with the category that is ordered last within the table.
Ordered categorical variables	Age	**Step 1**: If 0<n<11 for any cell within a category, then combine the youngest category for which this occurs with the next category (older age group) and update the range of the combined category. If n<11 in the oldest age group, then combine with the previous (younger) category. **Step 2**: Repeat the process described in Step 1 until all non-zero cells have n≥11.
De-identification of continuous variables	Age	**Step 1**: If standard deviation=0 for a treatment group (all values the same) then do not report any statistics for that treatment group (except for n). **Step 2**: If standard deviation>0, but n<11 for a treatment group, then do not report minimum, median, maximum, quartiles or percentiles for that treatment group. **Step 3**: If any summary statistics have been excluded during steps 1 and 2, add a footnote to the table stating, 'Summary statistics are not presented where they could lead to participant re-identification'.

EXAMPLES OF DE-IDENTIFICATION OF NON-ORDERED CATEGORICAL VARIABLES

Note all data included in the following examples is for illustration purposes only and is not based on any actual data.

EXAMPLE 1

	Treatment A (N=100)	Treatment B (N=100)
Race		
American Indian or Alaska Native	**2 (2%)**	15 (15%)
Asian	18 (18%)	15 (15%)
Native Hawaiian or Pacific Islander	20 (20%)	15 (15%)
Black or African American	15 (15%)	15 (15%)
White	15 (15%)	20 (20%)
More than one race	15 (15%)	18 (18%)
Unknown or Not Reported	15 (15%)	**2 (2%)**

Combine all categories that have at least 1 cell with n<11 (highlighted in red).

	Treatment A (N=100)	Treatment B (N=100)
Race		
Asian	18 (18%)	15 (15%)
Native Hawaiian or Pacific Islander	20 (20%)	15 (15%)
Black or African American	15 (15%)	15 (15%)
White	15 (15%)	20 (20%)
More than one race	15 (15%)	18 (18%)
De-identified	**17 (17%)**	**17 (17%)**

EXAMPLE 2

	Treatment A (N=100)	Treatment B (N=100)
Race		
American Indian or Alaska Native	**2 (2%)**	15 (15%)
Asian	18 (18%)	15 (15%)
Native Hawaiian or Pacific Islander	40 (40%)	26 (26%)
Black or African American	**4 (4%)**	**4 (4%)**
White	15 (15%)	20 (20%)
More than one race	19 (19%)	18 (18%)
Unknown or Not Reported	**2 (2%)**	**2 (2%)**

Combine all categories that have at least 1 cell with n<11 (highlighted in red).

	Treatment A (N=100)	Treatment B (N=100)
Race		
Asian	**18 (18%)**	**15 (15%)**
Native Hawaiian or Pacific Islander	40 (40%)	26 (26%)
White	15 (15%)	20 (20%)
More than one race	19 (19%)	18 (18%)
De-identified	**8 (8%)**	**21 (21%)**

De-identified category still has a cell with n<11 (highlighted in red). Combine this with category with lowest total n (highlighted in green).

	Treatment A (N=100)	Treatment B (N=100)
Race		
Native Hawaiian or Pacific Islander	40 (40%)	26 (26%)
White	15 (15%)	20 (20%)
More than one race	19 (19%)	18 (18%)
De-identified	**26 (26%)**	**36 (36%)**

Stop: All cells have n≥11

EXAMPLE 3

	Treatment A (N=10)	Treatment B (N=10)
Sex		
Male	**6 (60%)**	**5 (50%)**
Female	**4 (40%)**	**5 (50%)**

At least one treatment group has less than 10 participants, so all potentially de-identifying variables are de-identified.

	Treatment A (N=10)	Treatment B (N=10)
Sex		
De-identified	**10 (100%)**	**10 (100%)**

Stop: All necessary variables de-identified

EXAMPLE OF DE-IDENTIFICATION OF ORDERED CATEGORICAL VARIABLES

Note all data included in the following examples is for illustration purposes only and is not based on any actual data.

	Treatment A (N=100)	Treatment B (N=100)
Age Group		
0 - 27 days	**2 (2%)**	15 (15%)
28 days - 23 months	18 (18%)	15 (15%)
2 - 11 years	20 (20%)	15 (15%)
12 - 17 years	**6 (6%)**	15 (15%)
18 - 64 years	24 (24%)	30 (30%)
65 - 84 years	15 (15%)	**8 (8%)**
>= 85 years	15 (15%)	**2 (2%)**

Combine youngest category that has at least 1 cell with n<11 (highlighted in red) with the next age group (or previous age group for oldest category)

	Treatment A (N=100)	Treatment B (N=100)
Age Group		
0 days - 23 months	**20 (20%)**	**30 (30%)**
2 - 11 years	20 (20%)	15 (15%)
12 - 17 years	**6 (6%)**	15 (15%)
18 - 64 years	24 (24%)	30 (30%)
65 - 84 years	15 (15%)	**8 (8%)**
>= 85 years	15 (15%)	**2 (2%)**

Combine youngest category that has at least 1 cell with n<11 (highlighted in red) with the next age group (or previous age group for oldest category)

	Treatment A (N=100)	Treatment B (N=100)
Age Group		
0 days - 23 months	20 (20%)	30 (30%)
2 - 11 years	20 (20%)	15 (15%)
12 - 64 years	**30 (30%)**	**45 (45%)**
65 - 84 years	15 (15%)	**8 (8%)**
>= 85 years	15 (15%)	**2 (2%)**

Combine youngest category that has at least 1 cell with n<11 (highlighted in red) with the next age group (or previous age group for oldest category)

	Treatment A (N=100)	Treatment B (N=100)
Age Group		
0 days - 23 months	20 (20%)	30 (30%)
2 - 11 years	20 (20%)	15 (15%)
12 - 64 years	30 (30%)	45 (45%)
>= 65 years	**30 (30%)**	**10 (10%)**

Combine youngest category that has at least 1 cell with n<11 (highlighted in red) with the next age group (or previous age group for oldest category)

	Treatment A (N=100)	Treatment B (N=100)
Age Group		
0 days - 23 months	20 (20%)	30 (30%)
2 - 11 years	20 (20%)	15 (15%)
>= 12 years	**60 (60%)**	**55 (55%)**

Stop: All cells have n≥11

EXAMPLE OF DE-IDENTIFICATION OF CONTINUOUS VARIABLES

Note all data included in the following examples is for illustration purposes only and is not based on any actual data.

	Treatment A (N=10)	Treatment B (N=2)
Age (Years)		
n	**10**	**2**
mean	20.1	20
sd	12.23	**0**
min	7	20
median	19	20
max	28	20

Standard deviation=0 for Treatment B therefore remove summary statistics (except n).

	Treatment A (N=10)	Treatment B (N=2)
Age (Years)		
n	**10**	**2**
mean	20.1	.
sd	12.23	.
min	7	.
median	19	.
max	28	.

N<11 for Treatments A & B, so if not already removed, remove minimum, median and maximum (and any quartiles/percentiles if present). Add footnote if required.

	Treatment A (N=10)	Treatment B (N=2)
Age (Years)		
n	10	2
mean	20.1	.
sd	12.23	.
min	.	.
median	.	.
max	.	.

Note: Summary statistics are not presented where they could lead to participant re-identification.

CONCLUSION

Following the steps above will satisfy the requirement for de-identification of data prior to public disclosure and therefore lower the risk of re-identification. However, the approach should not be limited to the potential identifiers discussed in this paper. Consideration should be given to the potential for re-identification in other publicly available summaries, e.g., manuscripts, conference presentations or advisory committee materials, considering any variables that could potentially increase this risk of re-identification. For example, disease characteristics or other baseline / demographic characteristics could increase the potential for re-identification. Consideration is also given when presenting data for subgroups that could result in data displays with a small number of participants with unique characteristics.

REFERENCES

Samarati, Pierangela; Sweeney, Latanya (1998). *"Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression"* (PDF).

ACKNOWLEDGMENTS

Edwin van Stein, GSK for the creation of the examples.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Author Name: Nicola Perry
Company: GSK
Address: 79 New Oxford Street, London, WC1A 1DG
Email: nicola.x.perry@gsk.com
Website: www.gsk.com