Discussion: disclosure review and anonymisation

Uganda HIV study

Whilst full names have been replaced with initials/pseudonyms, the names remain in the file properties of the textual files.

Direct identifiers such as names have been pseudonymised / removed.

Possible indirect identifiers:

Variable	**Issue**	**Solution**	**problem**
Age at first interview	Actual age	Recode into age bands? e.g. 60-65	
Main activity	Cleaning/gardening area may to too specific	Reduce precision?	Loose information
Occupation	Job title quite detailed (Support staff at Entebbe Municipal Council; works at the Stone quarry)	Use more generalised descriptions, e.g. Public sector, Industry	Loose information
Household characteristics	Possibly identifying but likely not, e.g. says she lives alone in a rented house	Remove rented accommodation information?	Loose information
Location of interview	Not identifying: Uganda Virus Research Institute clinic, Nakawuka Health centre		
Location	Names of villages, districts etc where person has lived and/or lives now	Reduce precision?	

The life history information at the start of the interview narrative and the observational descriptions are extremely detailed, e.g. which tribe a person is from, place of birth, family size. Clearly, if a re-user of the data personally KNOWS someone in the study these specific attributes could result in that person being identified; or a re-user could trace the person in the area.

However, identifying someone we know in a study is NOT what we are evaluating when we undertake disclosure review, as this is highly unlikely.

Note the difference with SENSITIVE identifying information and variables that seem sensitive like HIV status/count/treatment regimen; whilst such information is sensitive, it is in itself not identifying in this dataset. The combination of information is the main issue to look out for in disclosure review. Identify and treat **identifying** variables or characteristics first.

Whilst the narratives and observational descriptions contain much potentially identifying information, they are also very rich in interesting contextual information, and have high re-use potential. Consider for example the value of this information in future as a historical record, describing an era of epidemic HIV infections across Africa; or its value to study research methods and the attitude of researchers towards participants.

In a dataset like this, consider the value of **access control** so that precision does not have to be reduced and so that is information not lost. Re-use value is increased if rich data is left. Historians rely on rich information from the past, so think about future value too.

Transcripts from the study were archived at the UK Data Archive. For the majority of interviews and observations only minimal anonymising was needed (name).

Crucial also in making this kind of data available to other researchers, is the fact that researchers re-using the data are bound by a legally binding end user licence which (amongst other conditions) requires a researcher to "*preserve at all times the confidentiality of information pertaining to individuals, households or organisations in the data files where the information is not in the public domain. Not to use the data to attempt to obtain or derive information relating specifically to an identifiable individual, household or organisation; and not to claim to have obtained or derived such information*."