Besides on the study’s context. however, there

Besides exploring the data, understanding it and its context is of vital importance; for setting a correct threshold when applying SDC methods the sample size and the sampling methodology must be known and understood. Direct identifiers are easily spotted when exploring the data; the most common direct identifiers are names and addresses/postcodes; open questions are most likely to contain direct identifiers and the more string variables containing answers of open questions the longer this step will require.

Steps 3, 4 and 5 involve the direct use of sdcMicro. About sdcMicrosdcMicro is an open source R package that facilitates the creation of anonymized microdata for secondary use. sdcMicro has been first released in 2007 and ever since further improved versions with more and better functionalities have been released constantly (a new improved Shiny graphical user interface has been made available lately as well; however this blog will discuss about the package and will include code snippets and output) . Choosing key variables (sample vs. population, k-anonymity and l-diversity)Choosing key variables can prove complex and time consuming and of course dependable on the study’s context. However, there are several methods that ease this process. Risk evaluation is centred on uniqueness in the sample and/or in the population; if frequencies are low in a large sample study it is highly expected for the frequencies to be low in the population – this is a first sign that a particular variable is disclose. To decide whether a case is or not a risk, a threshold is set; if the risk of a particular unit is above the set threshold than that unit is considered at risk.

There is no set threshold, as this depends on the level of access of a study, it’s proposed usage and types of users; however non-existence of a standardized threshold provides more flexibility to both data processors and data depositors. In order to get the individual risk of each case it is required to obtain the frequencies and acquire the key patterns in the population. The frequencies obtained are furthered used in obtaining two important estimates of individual risk: k-anonymity and l-diversity. K-anonymity is a statistical method to evaluate whether or not a case is unique in the data. Once again, depending on the level of access of a study the k is usually set to 3 or 5; so if k= 3 the cases that are unique in the data (f= 1) and the cases that have a duplicate (f= 2) will violate 2- and 3-anonymity assumptions and will be considered to pose a risk of disclosure.

The k-anonymity and sampling weights are used to obtain the ratios of the sample cases in the population; these are rough estimation of individual risk – not to be confused with the individual risk estimate. The individual risk estimates takes into account the distribution of the population frequencies given the sample frequencies, besides k-anonymity. Once the individual risk estimate is obtained for all cases an appropriate threshold (which can be deducted from the data based or agreed with the depositor) is set (e. g. 0. 05). Furthermore sdcMicro facilitates obtaining the l-diversity, an extension of k-anonymity, used to test whether the individual risk increases when a sensible variable is linked with the key variables (e. g.

BMI categories: underweight, normal BMI, overweight). If the l-diversity of each case is 1 than the sensible variable can be released as there is enough variability in the data and there is no possible disclosure risk created by this

variable. Special Uniques Detection Algorithm (SUDA/SUDA2)SUDA is used to obtain the disclosure risk for individual cases while considering a subset of the key variables. The risk estimation mainly depends on: sdcMicro provides further functionality for SUDA, of increased importance for a data provider when conducting disclosure review, the contribution (in %) of each variable to the risk. Having such a feature is extremely useful when deciding which variable to recode and how to recode them.

Cluster and Global RiskIncreasingly there are more and more data available that include cluster structures (such as the individuals of a household). sdcMicro provides the functionality of calculation cluster risk as if one particular case of a household has a higher risk this will increase the probability of disclosure for all the other cases in the household. Global risk estimates the overall risk of an entire dataset – this is an extremely useful feature as it provides a complete picture of the data at hand; however this must be used carefully, for example one method of calculating global risk – benchmark/threshold risk – might account for some cases that are or aren’t actually risky (falsely increasing/decreasing the global risk) as it depends on the distribution of all individual risk.