Guiding Principles For

Researchers seeking to utilise the Administrative Data Research Centre NI

1. Introduction
   1. As a partner in the Northern Ireland Administrative Data Research Centre (ADRC-NI), NISRA recognises that the opportunities of cross cutting, policy-relevant research are potentially far reaching. Such research can provide new insight into the things which matter the most to people in Northern Ireland, and indeed further afield, and through policy influence could help improve the key front line services in areas such as Health and Education that we all rely on so heavily.
   2. While the funding provided by the ESRC and the Public Health Authority has been pivotal to the success thus far, it has been matched by the commitment of those Government organisations who, recognising the potential for the ADRC-NI to be a game changer in maximising the value of the data they hold, have taken the positive step of making their de-identified data available for research purposes. For NISRA, maintaining trust and allaying any concerns that those organisations and indeed the public might have in terms of how their information is being used by the ADRC-NI is paramount. That critical obligation to safeguard both the privacy and the confidentiality of the information is one that we take very seriously, and that is precisely why we have put a comprehensive package of legislative, physical and operational safeguards in place. This includes, for example:

* ensuring that all of the information is de-identified prior to it being made available for research purposes;
* creating secure and supervised research facilities, which are only accessible to Approved Researchers to work on Approved Projects; and
* ensuring all outputs are suitably aggregated and non-disclosive.

1. Aim of the Document
   1. Against the above background, the aim of this document is to provide a number of guiding principles that researchers must adhere to when developing research proposals and the associated research application. Specifically, it sets out these principles in terms of (i) issues relating to data linkage, (ii) access to de-identify datasets for research purposes, (iii) definition of variables within the de-identified datasets deemed essential to the research and (iv) important considerations regarding access to unlinked data.
   2. At the outset, it is important to note that each research project will be considered on a case by case basis. When considering whether to engage with a particular project and indeed the extent of any engagement, Data Controllers will consider any risk to their current business activities and give due consideration to any potential sensitivities/ public perception issues that may arise through their involvement. Data Controllers may also consider the complexity of a proposed project from a resourcing perspective and be involved in agreeing priorities. In addition, they (along with the RSU) will give due consideration to the risk of inadvertent disclosure in (i) the finalised research dataset and (ii) the planned final outputs, the latter of which will be cleared by the RSU prior to publication. Data Controllers will also determine whether they require access to key research findings for the purposes of briefing, prior to release.
   3. In keeping with best practice, the guiding principles will be kept under review and amended by the RSU (taking account of the views of Data Controllers) as considered necessary, for example, in light of the Codes of Practice that will emerge from the planned Digital Economy Bill coming into effect.
2. Applying for access to linked datasets
   1. Researchers are advised that any data that they wish to link should be from a similar time period. There are risks associated with linking data from time periods significantly separated in time. The greater the time gap between the datasets to be linked, the greater the bias due to births, deaths, migration and name changes (e.g. through marriage). NISRA has undertaken research showing that this will have an impact on the quality of research and as a general rule, NISRA will only support projects with datasets separated by less than three years. Alternatively, research on static populations for example on people aged over 50 may be possible when the datasets are separated by a greater period of three years.
   2. As an example, when examining multiple Censuses, the researcher should use the Northern Ireland Longitudinal Survey (NILS) which incorporates health registration data at 6 month intervals to maintain high levels of data linkage.
3. Seeking access to de-identify datasets for research purposes
   1. The RSU, on behalf of the ADRC-NI and researchers, is responsible for securing access to the relevant data required by researchers for their project, through engaging with the relevant Data Controllers as necessary. Such negotiations will be undertaken on a project by project basis. Researchers should be aware that while the RSU will actively encourage engagement, it is not guaranteed - a Data Controller may decide, for a variety of reasons, not to support a particular project. In addition, where Data Controllers agree to support a project they may elect to impose restrictions on the records, variables and classifications within variables that are provided as they consider necessary. While such matters are outside the control of the RSU, the RSU will keep researchers informed about any restrictions and/or decisions that Data Controllers require/ make.
4. Defining the variables within the de-identified datasets
   1. When defining the variables that they require from a dataset, researchers should be mindful of, and actively consider, the following:

Developing a variable list:

* 1. Researchers should ask only for those variables and classifications within variables that are essential for their research – this helps to streamline subsequent project processes (e.g. disclosure assessments and the development and approval of the necessary data sharing agreement (DSA)) and experience has shown that applications with limited variables/classifications tend to be more favourably assessed by Data Controllers. This requires some level of thinking beforehand by researchers as to the purposes that each variable will be put and how it could relate to the particular research project in hand. Data Controllers and the RSU will also act robustly to ensure that a minimum dataset for specific pieces of research is created – this is in line with the new data protection regulations on data minimisation. Researchers are required to detail all of the variables that they require for their project in their application. However, recognising that omissions can happen from time, variables can in exceptional circumstances be added when the DSA is being developed provided sufficient justification is given.
  2. The [ADRC-NI Data Prospectus](https://adrn.ac.uk/get-data/catalogue/?Northern-Ireland) details the data sets which have been agreed in principle for use in the ADRC-NI. The [ADRC-NI Data Prospectus](https://adrn.ac.uk/get-data/catalogue/?Northern-Ireland) includes links to metadata where this is available but researchers can contact the RSU for details on the variables available in other datasets. As outlined earlier, Data Controllers will have the final say in terms of what data are provided from their particular dataset(s). Having said that, following the necessary disclosure checks, the IAO[[1]](#footnote-1) of the RSU will have the final say in terms of the information that is made accessible to researchers in the ADRC-NI Secure Environment. Such decisions will be taken in light of the totality of the data being provided rather than individual datasets and may result in, for example, the removal of certain variables and/or reducing the classifications within variables.

Variables in multiple datasets:

* 1. In the event where a Data Controller imposes a restriction on a particular variable or sub population within their data source, that restriction will extend to all other data sources being utilised in the project and will be reflected in the finalised de-identified data that are made accessible to researchers. For example, if a Data Controller requires that Country of Origin is aggregated or that dates are restricted to month & year in their particular data source, then those restrictions will be applied to, and pertain in, the finalised de-identified data to which they will be given access.

Justification for all variables:

* 1. Researchers must ensure that both the number of variables and their related classifications can be fully justified. Variables will not be provided for a project if, in addition to concerns regarding confidentiality and/or disclosure risk, the data provider considers that they are not essential for the research project in question, are inappropriate or are of insufficient quality/completeness.

Specific justification for restricted/sensitive variables:

* 1. Researchers must ensure that all restricted/sensitive variables, as identified by data controllers, are comprehensively justified for inclusion. For example, if detailed occupation is required the research would have to spell out why it is essential to their research and the implications for the research if it wasn’t available.

Variables with a high number of categories:

* 1. Access to variables with a large number of categories (more than 10 categories) or with individual categories that have a small number of records (less than 10 records) is likely to be restricted by data controllers. Researchers may be asked to consider how the data can be suitably aggregated and/or spell out why the variables and associated categories are essential to the research and the implications if they were not made available. Ultimately, Data Controllers will take the number of variables and level of detail into account when approving the list of variables and associated classifications and may replace detailed variables with less detailed versions. Experience has demonstrated that Data Controllers are much more comfortable being asked to release variables that don’t significantly increase the risk of re-identification.

Derived Variables:

* 1. Researchers will not have access to restricted/confidential or detailed variables. Data Controllers will process restricted/confidential or detailed variables to create derived variables. Researchers can request that NISRA RSU process restricted/confidential or detailed variables before use to create derived variables across several datasets.

The use of area-based variables:

* 1. Detailed geographical identifiers significantly increase the risk of re-identification and so researchers must consider the level of geographic breakdown required for their project. Data Controllers and/or the research network IAO may choose to provide only “anonymised” SOA variable (within LGDs) for ADRC-NI projects that clearly require such level of information. However, SOA based aggregated classification variables (deprivation deciles, rurality etc...) that don’t directly identify individual SOAs may be provided.  On occasion, NISRA RSU may also “de-anonymise” the SOA variable in significantly aggregated datasets for projects that have demonstrated a geography need – subject to approval by Data Controllers. A mappable dataset will be created by the researcher, approved by the RSU and provided separately from the main research dataset.  Also, because low geographical levels may lead to low numbers of observations, they will require special attention in relation to issues of disclosure control and confidentiality.

1. Applying for access to unlinked records
   1. There are added sensitivities and/or risks associated with unlinked records which Data controllers will factor into their decision making. Data controllers will consider requests for datasets where not all records are expected to link (e.g. linkage between a study population and a disease registry). However, access to unlinked records that have not linked for other reasons (e.g. missing information or poor data quality) will be more restricted by Data Controllers and access and the level of detail provided will be decided on a project by project basis.
   2. The researcher’s requirement to analyse the unlinked records should be comprehensively justified and Data Controllers may impose restrictions in terms of the variables and associated level of detail supplied. Geography variables and occupation/industry variables will not be included in either type of unlinked records dataset.
2. Further restrictions/adjustments
   1. Further restrictions adjustments may be made by the Data Controllers at the data extract stage prior to passing data to the TTP and the RSU after all the data have been reviewed. Details will be fed back to the researchers if possible within confidentiality constraints.
   2. Further restrictions adjustments may be made by the RSU after all the data have been reviewed. Details will be provided if possible.
3. Other references
   1. [ICO Code of Practice Anonymisation: managing data protection risk code of practice](https://ico.org.uk/media/1061/anonymisation-code.pdf)

1. The Information Asset Owners (IAOs) are the staff members in each business area who are responsible for overseeing the security of the organisation’s information and data. [↑](#footnote-ref-1)