NILS-RSU

Northern Ireland Longitudinal Study
Research Support Unit

Northern Ireland Longitudinal Study:

NILS Data Matching Methodology

NILS Working Paper

March 2025

NILS Research Support Unit

1. Introduction

The NILS is a large-scale record linkage study of a representative sample of approximately 28% of the Northern Ireland population that has been created by linking statistical and administrative data sources within Northern Ireland. These data include the 1981, 1991, 2001, 2011 and 2021 census returns, vital events (births, deaths and marriages), demographic and migration events, and property data. The NIMS is an additional major record linkage study that links the 2001, 2011 and 2021 census returns for the whole of the enumerated population to subsequently registered mortality data. For a detailed introduction to the NILS and NIMS, please refer to the NILS or Research Support Unit website.

This working paper provides a generic description of the technical methodology used to create linkages between each of the available NILS data sources.

2. Data Matching Objective

As with any data matching exercise, the objective is to join two datasets together based upon a common entity. The *'Spine'* of the NILS can be described as a sample of people registered with a General Practitioner in Northern Ireland. These people are considered to be the common entity to link between in each of the NILS data sources. To carry out a linkage between each dataset, certain information is used from each entity to compare and confirm/refute any potential linkage between datasets. The quality of matches between datasets is based upon the uniqueness of this information and similarity of data capture in each of the sources being linked. Typical information used across the NILS datasets includes name (forename(s), surname, previous surnames), sex, date of birth and address information.

Table 1 shows a list of the main NILS data sources and provides a description of how each of the datasets is populated.

Table 1: NILS Data Sources and Data Collection method

Data	Collection Description
Health Card Registration (Core NILS Data)	Data is collected using a combination of historical medical records and new information collected from GPs, hospitals, dentists and provided by patients. This data is referred to as Health Card Registrations.
Census (1981, 1991, 2001, 2011, 2021)	A member of the household completes a form on behalf of everyone within the household and then the details are scanned in using optical character recognition. In the 2011 and 2021 Censuses, an online form was also available for each household.
GRO Births	Information gathered from face-to-face interview with informant of birth – usually the mother or father of baby.
GRO Deaths	Information gathered from face-to-face interview with informant of death – usually a relative of the deceased.
GRO Marriages	Information gathered from face-to-face interview with informants of marriage.

The variety of ways that the data is collected and the amount of verification applied before the data is entered onto the systems invariably mean that the data will not be identical on all systems. The differences may be minor, such as a slightly different spelling of the surname or they could be significant such as an incorrect address, date of birth or a middle name used in place of a forename. Therefore, it is important to ensure that any matching algorithms take account of non-exact matched records.

3. NILS Data Matching Process

There is some variation in matching methodology between datasets due to individual characteristics of each data source. However, there is a general approach adopted when matching new data into NILS/NIMS for the first time. This takes the form of:

i. Data preparation;
ii. Automated data matching;
iii. Step 1 – Manual verification of automatic data matching;
iv. Step 2 – Multiple choice; and
v. Step 3 – Cross verification and Quality checks.

As previously mentioned, when matching data into the NILS, the reference or *'spine'* data source is the Health Card Registration data. For the NIMS, the reference data source is the corresponding Census from the same year: NIMS 2001 to all records in Census 2001; NIMS 2011 to Census 2011; NIMS 2021 to Census 2021.

i. Data preparation

The aim of the data preparation stage is to standardise corresponding match fields within the two data sources to be linked. At this stage the data is analysed to assess the key matching fields for quality and comparability to the reference data source. All identified issues are resolved where possible and two bespoke datasets are created for the automated matching phase.

Using the linkage of 2011 census data into the NILS as an example, it was identified that some name information was of lower quality (missing, mis-spelt etc) due to the Optical Character Recognition phase of scanning the 2011 forms. A clerical exercise was undertaken to correct these before matching proceeded.

Other anomalies include name information only including the persons' initials, postcodes in different formats, nicknames used rather than common names, day of birth transposed with month of birth etc.

ii. Automated data matching

Following the data preparation phase resulting in the creation of two bespoke datasets, an automated matching routine is created and run to produce potential linkages between the datasets. This process has two steps, Deterministic matching and FRIL matching.

Deterministic matching, referred to as Match-Keys, involves putting together pieces of information to create unique keys that can be used for automated matching. The variation in recording demographic information across datasets can occur in a number of different forms. A single match-key alone cannot resolve all of these differences, hence the need for multiple match-keys, each designed to resolve particular inconsistencies between match pairs. Table 2 shows a sample list of match-keys. Although the match-keys shown can be generically applied, the methodology is adapted for each project to meet the needs of differing datasets.

Table 2: Match-Key Example List

Match-Key	Description	Inconsistencies resolved by Match-Key
1	Forename, Surname, DOB, Sex, Postcode	None- exact agreement
2	Forename Initial, DOB, Sex, Postcode	Name discrepancies
3	Surname Initial, DOB, Sex, Postcode	Name discrepancies
4	Forename, Surname, Age, Sex, Postcode	DOB discrepancies
5	Forename, Surname, DOB, Sex	Movers out of area

The match-keys are processed in a stepwise manner starting with match-key 1 and working down to the last match-key. Records are only linked on a match-key if it is unique on both datasets (i.e. one-to-one match). If multiple records match on a particular match-key then the link is not made and candidates are passed on as a residual to the next match-key.

The second of our automated processes utilises a dedicated record linkage software package called Fine-Grained Records Integration and Linkage Tool (FRIL)[1].

During this process FRIL implements a number of data linkage cycles between datasets using a different configuration at each stage taking account of variation in the fields used.

A match confidence score is assigned to each record pair based on the strength of similarity between matching fields. The number of records processed is reduced during each matching cycle with records associated with strong matches removed before the next cycle runs.

As a result of the automated matching phase the data can be divided into four sections:

a) One-to-one data matches having a match score identified as being over a pre-determined threshold;
b) Data matches where the match score is above the threshold mentioned in a) but linked to more than one person;
c) One-to-one matches with a match score below the threshold; and
d) Unmatched records.

[1] http://fril.sourceforge.net/

iii. Step 1 – Manual verification of links

Following the automated matching process, a random sample of the one-to-one data matches achieved from deterministic matching and those identified through FRIL as having a match score above the threshold are manually verified by matching staff. The objective of this phase is to affirm the quality of matches produced in the automated matching phase.

At this stage, matching staff visually compare information from the two linked datasets using bespoke Graphical User Interfaces (GUIs). The matching staff have the ability to take additional information into account (names, addresses, age, sex, other household members etc) and confirm/refute any automatically generated matches. If a considerable number of rejected links are identified then further *'tuning'* work may be carried out on phase ii and the automated matching re-run until satisfactory matching levels are reached.

iv. Step 2 – Multiple choice

The fourth phase of NILS data matching involves looking at two types of data matches generated from the automated matching phase:

- Data matches where the match score is above the pre-determined threshold mentioned in a) but linked to more than one person; and
- one-to-one matches with a match score below the threshold.

Again, matching staff use specifically designed GUIs to look at information from the potentially linked datasets and to confirm/refute the generated matches.

v. Step 3 – Cross Verification and Quality checks

On completion of the matching, all links will have been created, and a further manual verification exercise is carried out on the links to identify errors and quality. An analysis is carried out on the linkages and other information to identify any inconsistencies between datasets. Linkages associated with any discrepancies are marked up for manual verification and checked in the same manner as described in phase iii of the matching process.

Some examples of the analysis include:

- *Births to NILS members* – number of linkages for the mother exceed the total of previous births as noted in the GRO births file, dates between births appear incorrect as they are too close together, DOB of the baby is different on the Health data and the GRO data; and
- *Deaths* – discrepancies between address and registration district, health data not marked as NILS member being deceased, differences in the person's name and different date of birth on the Health data and the GRO data; and

- *Marriages* – persons getting married are younger than 16, persons getting married more than once in a 12 month period, health data marked as NILS member being deceased prior to the marriage taking place.

A random sample of matches similar to those described in phase iii are also included for verification as a quality check on the entire matching process.

The results of Step 3 – Cross Verification and Quality Checks are analysed for consistency and if no issues are identified then the matching work is marked as complete. If any potential issues are identified then there is remedial work carried out to resolve the issues, which may include more manual verification, before the matching is marked as complete.

4. Results

This matching method has been used to create the links in the NILS and NIMS datasets. Table 3 shows the number of matches for NILS and NIMS linkages involving census records. An overall adjusted match rate is given to account for BSO list inflation, census imputation and/or census enumeration rates, further details are available in NILS Working Paper 4.0.

Table 3: NILS/NIMS Census Match Rates

	Type of Linkage	Records Linked	Adjusted Match Rate
NILS	Health Card Registrations to Census 2001	457,470	99.9%
	Health Card Registrations to Census 2011	486,711	99.2%
	Health Card Registrations to Census 2021	516,042	96.6%
NIMS	Census 2001 to deaths	135,900	94.1%
	Census 2011 to deaths	182,128	96.5%
	Census 2021 to deaths	42,426	95.3%

BSO health data list inflation can be caused by population movement without GP knowledge e.g. people emigrating, migrants returning home and deaths of residents outside the jurisdiction. Another identified causation is non-entitled users such as Republic of Ireland residents accessing Northern Ireland health services.

Since Census 2001, comprehensive processes have been utilised to adjust for over coverage and under-coverage in the enumerated population, to produce an accurate estimate of the whole Northern Ireland population on Census Day. Some of these records are unsuitable for matching to the NILS data due to their source, as they do not have any of the key demographic information to enable matching. Table 4 shows the adjustment rates for list inflation and under-coverage adjustment rates for each of the censuses.

Table 4: Adjustment rates for list inflation and census under-coverage adjustment

Census	List Inflation	Under-coverage Adjustment
2021	5.2%	1.7%
2011	4.1%	4.8%
2001	4.7%	4.6%
1991	4.7%	N/A

Table 5 shows the number of matches for all NILS health card registrations linked to GRO vital events and a typical match rate (using a 10 year average) that has been achieved using this methodology.

Table 5: NILS Match Rates for GRO Vital Events

Type of Linkage	Records Linked	Raw Match Rate (10 year average)	Years of Linkage
Deaths of NILS Members	135,640	99.7%	1991-2022
Births of NILS Members (babies)	332,006	90.4%	1974-2022
Births to NILS Fathers	284,410	97.1%	1974-2022
Births to NILS Mothers	319,613	99.0%	1974-2022
Marriages (female) of NILS Members	38,719	93.2%	2005-2022
Marriages (male) of NILS Members	36,735	89.4%	2005-2022
Widowerhoods of NILS Members	48,351	86.7%	1991-2023

The raw match rates for NILS health card registrations linked to GRO vital events will be affected by several factors such as BSO list inflation and non-resident vital events registrations. Not all births, deaths or marriages will be expected to link to the NILS data as a number of those registered may be for those who are not resident within Northern Ireland.

5. Conclusion

Considerable research and experience have gone into the development of the NILS matching process. This has resulted in an optimum matching methodology that generally yields a high match rate when linking NILS and NIMS data.

Matching methods are continually changing and developments in technology enable more efficient matching to be implemented.

NISRA

March 2025