Imperial College London

NDAU
Neonatal Data Analysis Unit

Neonatal Data Set v2.0 Data Quality Assessment

1 Overview

1.1 Summary

The purpose of this document is to provide details of the current data quality assurances applied to the Neonatal Data Set (NDS) in order to create the National Neonatal Research Database (NNRD), a National Data Asset designed to support multiple outputs including commissioning, service evaluation, audit, surveillance, quality improvement, policy development and research, fulfil the principle "*Collect once, use for multiple purposes*" and eliminate the burden and waste of duplicate data collection. Further to this, future improvements to the current pipeline are outlined.

2 Data quality processes

Currently a defined data extract, comprising the variables that make up the NDS (ISB1595) is received quarterly by the Neonatal Data Analysis Unit (NDAU) from Clevermed. Clevermed are the authorised NHS hosting company and supplier of the BadgerNet electronic patient record system.

Clevermed, apply the 'National Opt Out' before data are sent to the NDAU.

Data extracts are provided by Clevermed in the form of password protected ".**bak file**"; a ".**bak file**" is a file used by database applications such as SQL Server, to signify a backup copy of a database. Clevermed upload a link to this .bak file and share this with NDAU over the N3 environment. The .bak file received contains 10 tables, each table reflecting the multiple data entry screens that comprise the BadgerNet neonatal patient management system. The tables cover the following categories of data:

- The Episodes table represents admissions to a neonatal unit; this table contains personal identifiers, admission details, baby demographics and birth details
- The Daily data table corresponding to all daily treatments and care processes while a baby is on the neonatal unit
- The Diagnoses table comprises all procedures and diagnoses a baby receives while in the neonatal unit
- The rest of the tables cover results of retinopathy screening, abdominal x-rays, cranial ultrasound scans, microbiological cultures, two year follow up health status, hearing test and biochemical screening; collectively these tables are referred to as the "ad-hoc tables"

The .bak file is downloaded onto an NHS computer at the Neonatal Data Analysis Unit; the data in the file is then restored onto a secure NHS server. The newly restored database contains a number of identifiers and identifier codes assigned by Clevermed:

- an "Anon Patient ID" reflecting each episode of care; thus a baby that is transferred between neonatal units will have multiple "Anon Patient ID" codes (this identifier has replaced the previous baby specific Badger3 Anon ID)
- an episodic "Entity ID" that links a baby to each admission
- an encrypted NHS number, "National ID Baby Anon"

- a Lower Super Output Area (LSOA) code which reflects a geographical location for each postcode in England and Wales; this is reassigned using the most up-to-date information from NHS Business definitions

Section 2.1 describes how current data quality checks are implemented to create the NNRD from quarterly NDS extractions.

2.1 Current processes

The data quality checks outlined below represent those currently employed.

2.1.1 De-identification

- Under the current data flow, personal identifiers such as NHS number of baby and mother; mother's postcode; date of birth of mother and baby; baby admission and discharge dates) are removed from the Episodes table and are held in a separate location on a secure NHS server accessible only by the designated Senior Analyst. A baby can be linked to the identifier file using the following identifiers:
    - Anon Patient ID
    - Encrypted NHS number
- The Daily, Diagnoses, and ad-hoc tables do not contain any personal identifiers, only the "Entity ID" and "National ID Baby Anon" remain; the date and time of data entry to these tables are removed.

2.1.2 Checking provider codes

- Each episode of care record in the Episode table has a "Provider NHS Code"; these are checked against a pre-existing list of providers (the "Master List") to ensure all the providers shown in the in the file can be identified; if a new "Provider NHS Code" is found, Clevermed is asked to confirm the neonatal unit represented by the code; this is added to a Master List for future reference
- The Master List is used to identify the hospital/neonatal unit providing care, the Operational Delivery Network to which they belong and the designation of the neonatal unit (NICU, LNU, SCBU)
- This process is also carried out on "Place of Birth NHS Code", "Booking NHS Code", "Admit From NHS Code" and "Discharge Hospital NHS Code"

2.1.3 Minimum data entry check

- An initial minimum data quality check is carried out on all records in the Episodes table to identify records with no birth year, gestational age, admission time or negative episode number; these records are removed and held in a separate table

2.1.4 Addition of items required for audit reporting and merging

- Episodic data for each baby is compared to find the last known episode, and the last known neonatal discharge; if these details indicate the end of neonatal care the baby is assigned a "FinalYear" and "FinalQuarter" to indicate discharge from neonatal care

- The date of the extract is included in the dataset under the field 'File origin'. This tells the analyst from which extract the records originate and is used in the merging step detailed below

2.1.5 Removal of episode duplicates

- Episode duplicates are removed in two ways:
    - Any records with a duplicated "National ID Baby Anon" identifier and "Episode Number" are taken out of the NNRD and stored in a separate table
    - Any records with a duplicated "National ID Baby Anon" identifier, "Admission Time" and "Provider NHS Code" are taken out of the NNRD and stored in a separate table

2.1.6 Finalising the Daily, Diagnoses and ad-hoc tables

- Once the Episode table has been checked, the "Entity ID" of the records in it are used to extract any relevant data from the Daily, Diagnoses and ad-hoc tables with a matching "Entity ID". This ensures all data from an eligible admission are kept
- Any duplicate records in the Daily, Diagnoses or ad-hoc tables are removed. A record would be considered a duplicate if other records with the same compound key exists within the same table

2.1.7 Merging the new extract into the pre-existing data

- Once the tables in the newest extract have been cleaned, they are merged into the pre-existing NNRD (data from 2007 onwards)
- "Entity ID" is used to identify which records on the Episodes table in the new extract are also in the pre-existing NNRD Episodes table; where there is overlap, these records are deleted from the pre-existing Episodes table and are replaced with the records in the new extract; this allows any changes made to the data between quarters to be incorporated into the NNRD
- Any new records that appear in the newest Episode table but not in the pre-existing Episode table are also added allowing data from the most recent quarter to be added the NNRD
- The same process is carried out on the Daily, Diagnoses and ad-hoc tables so that any older records are removed and replaced with their newest versions and any new records are incorporated into the NNRD
- Once merging is complete, another round of duplicate checks are applied to all the tables in the merged NNRD to be sure none are introduced during the merge step
- Where any changes to the structure or coding of items in the EPR system have occurred between the previous download and the next, these must be accounted for before merging occurs

2.1.8 Quality assurance

- At present, each quarter a Quality Assurance report is sent to the clinical lead for each neonatal unit; the report contains information on data completeness, potentially erroneous entries and the relevant Badger IDs so that clinical teams have opportunity to check their

entries; any changes or corrections are entered into the real-time Clevermed platform and thus received by the NDAU at the next quarterly download

- Hospitals in England and Wales that submit data to the NNRD are included in this reporting. Restrictions to the processing of Scottish data mean that currently no quality assurance reports are produced for Scottish units
- Data from neonatal units relating to the current quarter of admission form the basis of the Quality Assurance reports
- The NDAU Manager and an analyst are available to respond to queries from clinicians

2.2 Plans for future development

In this section, plans for further improvements to data integrity and utility are described.

2.2.1 More frequent and complete data extracts

Moving towards a monthly data extract would mean reporting would be more up to date and reflect current practices. In light of the recent pandemic this has obvious benefits.

In addition to increased frequency, each extract would then reflect the entire dataset and avoid the need for merging which can introduce identifier clashes. In addition, NHS numbers should be cross-checked against demographic details held by the NHSD Personal Demographic Service.

2.2.2 Immediate feedback and quality reports for providers

It would improve quality to give clinical teams timely feedback on all or key NDS items.

More timely feedback would allow more complicated data quality issues to be resolved; these include for example:

- Internal inconsistencies, where items contradict one another
- Temporal inconsistencies, where timings of events do not make sense or appear to be erroneous in the context of other entries
- Out-of-range items
- Internal illogic where items are impossible (such as a very preterm baby having a very high birth weight)

2.2.3 Sign-off

Sign-off by the UK Neonatal Collaborative lead, a designated member of the neonatal clinical team for each NHS neonatal unit

2.2.4 Drug entries

These are currently inconsistently recorded (spelling, generic name, trade name, popular name); we plan to recode to produce a consistent entry in the NNRD

2.2.6 Changes to EPR systems to be documented by the Supplier and shared promptly

Where changes are made between downloads, these should be documented by the supplier and these details should be shared with NHS Digital.

2.2.7 Data interoperability

NDS data items are currently being mapped to the Observational Medical Outcomes Partnership *Common Data Model* (*OMOP-CDM*) and SNOMED-CT; this work will be completed by spring 2022.

2.2.8 Data linkage

The NNRD will be linked within a secure Trusted Research Environment to other health and non-health data sources to widen utility and support the UK Government "*Levelling-Up*" agenda

2.2.9 Trusted Research Environment

Funding has been secured to move the NNRD to be located within a secure NHS Trusted Research Environment; this work is scheduled for completion by autumn 2022.