necs Partners in improving local health

NHS
North of England
Commissioning Support Unit

Open Safely Processing Methodology

Background

In support of the, The OPENSafely project, High Cost Drugs patient level data submissions for all of 2018/19 and 2019/20 were requested from all regional Data Services for Commissioners Regional Offices (DSCROs). Around 150 data submissions in approx. 20 different formats were received covering data from all Acute Providers and Commissioners nationally. We are unable to say with any certainty that the data received is a complete data set for each Provider and Commissioner.

The data was loaded and standardised into an agreed specification. The data was pseudonymised and then linked to the TPP GP Clinical Record data as provided by TPP, the subset of linked data was then shared with OPENSafely. Linkage only occurred where a valid NHS number was submitted. If an NHS Number was not provided or was invalid linkage will not have occurred or could have linked incorrectly. No secondary linkage approach was utilised due to the data quality.

Data has been received in around 20 separate schema formats, and reasonable best endeavours have been taken to map, load and standardise the data, though processing as documented has not been exhaustive, and further processing, cleansing and validation work could be applied to the information if required. Work undertaken nationally to review and improve the data quality of drugs submissions will vary greatly at across different regions. In a lot of instances it will be focused on key fields that are utilised operationally. Newer fields such as SNOMED code related fields are far less reliable as clinical systems to enable the population of such fields improve.

A standard data validation process could not be implemented due to this being a one off data collection exercise. For example, normally the process would compare the data submission to the previous month's submission, to ensure the growth is in line with expected trajectory and to ensure no loss of expected or inclusion of unexpected data items. NECS are unable to comment on the level of validation that has been undertaken on the data before it was received. The validation and data cleansing that we were able to undertake is as detailed in this document.

NECS therefore recommend that any assumptions or outcomes, based on this data, should be appropriately caveated.

This document details this process in more detail.

Open Safely Base Processing and Storage

- Data was loaded and archived as Raw Text rows and held by the DSCRO
- Around 150 files in around 20 different non-standard schemas were loaded
- Data was scanned for legally restricted codes to prevent unwanted disclosure of sensitive details
- Remaining PCD was then pseudonymised with Age replacing DOB and the agreed pseudo key applied to NHS Number
- Data was then typed into Numeric and Date fields as per the specification
- A number of fields were requested in the data submission that was excluded from the final extract provided to OpenSafely. This included fields such as Provider and Commissioner code, that were needed to be able to sense check the data received but that were not relevant to the final use of the data.

No patient identifiable data was shared as part of the final specification and extract.

Field Derivation

Due to the number of different data specifications that were received, a 'one size fits all' approach was not possible. The available data was utilised to try and populate some of the gaps in the data. The following approach was taken:

- Some of the submitted data included age. Where this was provided, the submitted value has been mapped straight into that field. Where age was not provided, but Clinical Intervention Date was present; the clinical intervention date was used along with DOB to derive an Age where possible.

- We have derived SNOMED code from drug name where possible matching on the description.
- We have also looked up the VTM (virtual therapeutic moiety) code and description for each drug based on Drug Name (Fully Specified Name) and SNOMED Synonyms where no match found on full name.
- The derived SNOMED information is included as additional fields at the end of the specification with a field name pre-fixed 'derived….'. The other SNOMED fields are as submitted by the Provider.

Additional Data Cleansing Measures

The data quality of the submitted data varied greatly. In a lot of instances the fields were populated with meaningless values that could not be utilised to provide anything more significant. In these instances the values have largely been left as submitted. The following data cleansing steps were taken:

- Data is scanned for duplicate rows based on NHS Number, Provider Code, Commissioner Code, Gender Code, Treatment Function Code, Drug Name or Drug Code, Point of Delivery, Date of Delivery, and Cost. Where all these relevant values are equal, the duplicate records were removed and a single record added back to the data set
- The national drugs data submission can include none activity PODS such as adjustments for example. POD field was requested in the data collection, and any none activity PODS were excluded from the final OpenSafely extact that was provided. Elements of this none activity data may still be included in the final extract if the data was not coded correctly.
- Small elements of pre-1819 was received in the data collection, this has been excluded. Future months in 2021 have been included and low volumes of wider 2021 data are present in the output.
- Gender Code was cleansed, where values can be confidently mapped to the intended standard i.e. where Value = 'f' then the value is updated to = 2,
- Year was cleansed to standardised format where it could be confidently done (for example 1819, 201819 etc)
- Erroneous Withheld ID values were removed
- Missing month values were populated from 'drug administered date' where applicable
- Treatment Function Codes were restricted to national values only, any non-national values were set to null.

Precise methodology of additional cleansing steps below:

StoredProcedure 2.sql

Data has been filtered to exclude patients not requested in the extract and to exclude records whose purpose is solely financial and does not relate to drugs in a clinical context (for example, adjustments).

Summary Numbers:

- Total Record Count in source data: 20,127,617
- Total unique Provider codes Count* in source data: 174 (167 validated against national reference data) variance of 7*
- Total unique Commissioner codes Count* in source data: 212 (179 validated against national reference data) variance of 33*
- Total unique Pseudo NHS Numbers: 999066

*This could include providers and commissioners that have subsequently merged, it could also include historic and new commissioner codes where they have changed in year. Some rogue codes may also be present accounting for the un-validated codes, due to data quality. Variances above show where non-national Provider or Commissioner codes were submitted.

Output Details:

- ¬ Delimited
- No Text Qualifier
- Additional Fields have been added where NECS derived SNOMED code from drug description. We did not merge these with the existing fields in order to aid analysis and data quality assessment. The fields are at the end of the data specification, contain text values and are entitled:
    - [Derived_SNOMED_From_Name]
    - [Derived_VTM]
    - [Derived_VTM_Name]

Notes

Data was received from 174 Providers and 212 Commissioners. NECS has no way of verifying if the data received for each Provider and Commissioner is a complete position and reflective of all activity.

TPP SystmOne representatives provided NECS with a list of Pseudo NHS Numbers taken from the GP Clinical system. These were used to identify the subset of data that has been shared onwards.

Data sharing

The pseudonomised output was compressed and password protected using 7-zip software and issued to NECS DSCRO to securely transmit the file back to TPP, using NHS Digital MESH routine. The password was then forwarded to Johnny Cockburn, this was relayed via SMS message as requested.

The information governance for this work was agreed and managed between NHS Digital colleagues and OPENSafely representatives. No data was shared until both parties gave the green light to proceed on the basis of all information governance protocols being in place.