NPDA 2023-24 Report – Notes on Data Validation and Cleaning

Introduction

This document summaries the processes involved in the validation of NPDA data, from pre-submission through to the final quality assurance.

Data Validation and Cleaning Process

Stage	Process
Validation prior to submission	• Data collection methods and dataset changes are piloted and refined. • Data quality warnings and data completeness report generated upon each upload enable the summitting unit to identify gaps and formatting errors in their data. • Clinical lead signs form to confirm that they have reviewed their data, that no data has been submitted on behalf of patients wishing to withdraw, and that they confirm that the data are accurate. • E-mail merge to clinical leads asking to confirm name of clinical lead, unit, Trust, region. • Download most up-to-date reference files used in the analysis.
Validation of data downloaded from system	• The list of units within the dataset is compared against the list of units with clinical lead sign-off form. • 10 random unit level csv files are downloaded, and data are compared with the data in the extract file to make sure that the data submitted by the unit has been correctly extracted from the data capture system. • The number of patients submitted by each unit are compared against the numbers submitted in the previous year and units with a difference of above or below 15% are contacted. • Data in all columns are checked to make sure data are in the correct columns.

	• Data quality of key variables are checked at unit level e.g. date of birth, date of diabetes diagnosis, gender, ethnicity, diabetes type, postcode and HbA1c. Units are contacted when more than 50% of observations in any key variable is unknown, invalid, or missing.
Building of master file for cleaning	• Duplicate rows are identified (i.e. rows where every observation in the corresponding row are an exact match) and only unique rows are kept. • Blank and erroneous rows are deleted (i.e. rows where all observations are blank or contain information that is not NPDA data). • Record id assigned to each visit. • Questionnaire data merged with extract file. - Check to see which units submitted using a mixed questionnaire/csv upload method. - Check that units employing a mixed method approach only have questionnaire data entered subsequent to csv data entered, as new csv files should overwrite data entered via questionnaire. - If any questionnaire data is found, escalate to project manager and identify whether it should be incorporated or deleted.
Data cleaning and preparation for analysis	• Variables formatted correctly e.g. eliminate typos, spaces, comments, make sure dates are in the correct format. • Keep valid observations only e.g. eliminate records outside of the audit period, values outside range, codes not matching the dataset specification, and illogical records like date of diabetes diagnosis before date of birth. • Individual characteristics that don't change over time are homogenized by children e.g. date of birth, date of diabetes diagnosis, sex and ethnicity. If some of these characteristics are unknown or missing within the current dataset, they will be searched in previous years datasets. • Add in data from reference files: - LSOA - Index of multiple deprivation - GP practice code - ICS - NHS region - Regional Network
Validation of statistical models and algorithms	• A separate analysis of health check, outcome and case-mix data is completed by the statistician in a separate statistical package. • Statistician, analyst and project manager review both analyses, discuss any inconsistencies and agree final analysis.

Final quality assurance	• A separate aggregation of regional and unit level data is undertaken and compare against the final aggregated data prepared by the analyst.