



Validation of National Child Measurement Programme data

This paper sets out the principle and rules which will be used to validate the National Child Measurement Programme data (NCMP) which is collected via the NCMP online system.

Published November 2016

Contents

Introduction	3
NCMP Validation Principle	3
Advantages	3
Disadvantages	4
At Source Validations	5
Data Quality Table	9
Post Deadline Validations	10
What does NHS Digital do when a data quality issue is identified?	10
Data Quality Table	10
Additional post-deadline validations	12
Annex A: Calculation of extreme values	13

Introduction

The principle explains the rationale behind the rules which follow. The rules fall into two groups:

1. At source validations – these fall into two groups:
 - a. Errors - these stop a record from being allowed into the system as they would invalidate the dataset, e.g. the same child details being entered twice.
 - b. Warnings – these ask local authorities (LAs) to confirm the record is correct as the system has identified it as a potential error, e.g. extremely high weight measurements.
2. Post deadline validations – these validations are carried out by NHS Digital once data from all LAs has been submitted and the collection period has closed. Any queries are then raised with the LA concerned.

We welcome feedback from users of the NCMP online system on any additional validations they think we should consider. These should be sent to enquiries@nhsdigital.nhs.uk using “NCMP validation suggestions” as the subject heading for the email.

NCMP Validation Principle

The main principle behind the validation of NCMP data is to push the onus for data quality onto the data provider.

This then encourages the provider to take ownership for the quality of their data and avoids costly workarounds to clean data centrally.

Therefore all data which the providers enter onto the NCMP system will be used even if there is concern around the data quality of some records.

These data quality issues will be dealt with in two ways:

- i. Adding a caution flag to the underlying dataset so other users of the data can make a decision on whether to include these data items in their analyses.
- ii. Using the data quality note which will accompany the national report to make users aware of data quality issues.

The only time when a record will be excluded from **the analysis carried out by NHS Digital** will be when it is adversely impacting on national level results.

This principle is in line with other data collections within NHS Digital.

Advantages

1. Pushes the onus for data quality onto the data provider.
2. Naming providers who appear to have poor quality data should encourage them to improve the quality of their data in future submissions.
3. Reduces resource needed for validation centrally and allows for earlier publication.
4. Avoids the potential for conflict if a record is removed from the national report that the data provider insists is correct.
5. Alerts users of the data to data quality issues and allows them to make an informed decision on whether to include or exclude specific records.

Disadvantages

1. Includes some data in the dataset which appear to be outliers.
 - a. However, this will be mitigated against by adding data quality caution flags to the underlying data.

At Source Validations

The NCMP system validates data as it is entered. For each record the system checks that all mandatory fields have been populated and that each field contains valid data. Records with missing mandatory fields or invalid data cannot be saved (classified as “rejections”). Records with improbable fields (e.g. extreme measurements) can be saved but will generate “warnings” that the data provider will need to confirm before they can finalise their submission.

Data can be entered into the NCMP system in two ways and validation can differ depending on the approach used. The two ways of entering data follow along with information on how rejections and warnings are handled:

1. **Individual record data entry** – data are typed directly into the system by users:
 - a. Rejections: it will not be possible to save the record unless all entered data are valid and mandatory fields have been populated.
 - b. Warnings: it will be possible to save records with warnings but all warnings must be confirmed by the data provider before the submission can be finalised
2. **Multiple record data entry** – data are uploaded into the system via a CSV file:
 - a. Rejections: records containing invalid data and/or missing mandatory fields will be flagged to users for correction before their upload can be completed.
 - b. Warnings: records with warnings will be flagged to users and can be confirmed or corrected before the upload is completed, or they can be uploaded with the warnings. All warnings must be confirmed by the data provider before the submission can be finalised.

The following validations are performed by the NCMP system:

1. **URN (DfE school number) – Rejection** - This is a mandatory field and must be populated with a valid URN.
 - Individual data entry: data item is automatically populated by the system based on the school grid the user is adding data to.
 - Uploaded data: URN must belong to one of the schools on the user’s schools list.
2. **Date of Birth (DoB) – Rejection** – This is a mandatory field and must be populated with a valid DoB. Valid DoB ranges are set each year as part of collection year set up. These ranges ensure that only children aged 4, 5, 10 or 11 years of age can be entered into the system. As an example, the ranges set for the 2014/15 collection year were: 01/09/2009 to 01/09/2010 for Reception and 01/09/2003 to 01/09/2004 for Year 6.
 - Individual data entry: the user must select a valid DoB from dropdown menu.
 - Uploaded data: each record must contain a valid DoB in dd/mm/yyyy format.
3. **Sex – Rejection** – This is a mandatory field and must be populated.
 - Individual data entry: the user must select a sex from dropdown menu.
 - Uploaded data: each record must contain one of the following valid responses: M, F, m, f, MALE, FEMALE, male, female.

4. **NHS number – Rejection** – This is not a mandatory field and can be blank. However, users are strongly encouraged to provide this data item since it will allow individual NCMP records to be linked such as a child's reception and year 6 measurements. The proportion of records with a missing NHS number is shown in the data quality table (see later). Where a value is provided, the system checks that the NHS number is valid using the Modulus 11 algorithm as detailed here: http://www.datadictionary.nhs.uk/version2/data_dictionary/data_field_notes/n/nhs_number_de.asp?shownav=0.
5. **Ethnicity – Rejection** - This is not a mandatory field and can be blank although users are strongly encouraged to provide this data item since it is an important variable for analytical purposes into inequalities. The proportion of records with a missing ethnicity code is shown in the data quality table (see later). Where a code is provided it must be included in the codes list held in the NCMP system. Currently the NCMP System accepts 342 ethnic codes taken from the following sets of ethnicity codes: four-character DfE ethnicity codes; single-character NHS ethnicity codes; RIO ethnicity codes and SystmOne ethnicity codes. These are shown in the reference data tables on the NCMP system homepage: http://content.digital.nhs.uk/media/18419/NCMPReferencedata/xls/NCMP_reference_data.xlsx

6. Height and weight – There are two checks carried out:

- **Rejection** – “error” ranges have been set for height, weight and BMI. If an entered height or weight falls outside the relevant range, or their combination produces a BMI that falls outside the BMI range, then the record will be rejected. The ranges have been set by analysing historical NCMP data. The following ranges are currently in use:

Measurement	School Year	Min	Max
Height	Reception	70 cm	150 cm
	Year 6	90 cm	200 cm
Weight	Reception	7 kg	70 kg
	Year 6	10 kg	160 kg
BMI	Reception	9 kg/m ²	45 kg/m ²
	Year 6	9 kg/m ²	70 kg/m ²

- **Warning** – “warning” ranges have been set for height, weight and BMI. These ranges use standardised values¹ (z-scores) which take child age and sex into account. Measurements resulting in z-scores outside the “warning” ranges will generate warning flags. It is possible to save records with warning flags but all warnings must be confirmed by the data provider before the submission can be finalised. The “warning” ranges currently in use are z-scores lower than -3 to or higher than +4.

7. Date of Measurement (DoM) – There are two checks carried out:

- **Rejection** – Each year a DoM range is set as part of collection year set up. If an entered DoM falls outside the range then the record will be rejected. This ensures that the child has been measured within the correct academic year. For example, the range used for the 2014/15 collection year was 01/09/2014 to 31/08/2015.
- **Warning** – If an entered DoM is a weekend day or in the month of August then a warning flag will be added to the record since schools are presumed to be closed on these days. It is possible to save records with warning flags but all warnings must be confirmed by the data provider before the submission can be finalised.
- Individual data entry: the user must select a valid DoM from the dropdown menu.
- Uploaded data: each record must contain a valid DoM in dd/mm/yyyy format.

8. Child postcode – Three checks are carried out on the child postcode:

- **Rejection** – The postcode must be in a valid format. It is important to note that this check is just against the format of the postcode and not whether the postcode is actually valid as doing so would make the NCMP system run too slowly. For example, LS99 9ZZ would pass the format check but it is not a

¹ The standardised value is called a z-score and indicates how far, and in what direction, the measurement deviates from the average (mean) for that age and sex (see Annex A for more detail).

valid postcode. Therefore, all postcodes should be checked for validity before they are entered into the system. The following postcode formats are accepted by the system: A9 9ZZ; A99 9ZZ; AB9 9ZZ; AB99 9ZZ; A9C 9ZZ; AD9E 9ZZ².

- **Warning** – Child postcode is not a mandatory field and can be blank although users are strongly encouraged to provide this data item since it is required for analyses based on child residence. The proportion of records with a missing postcode is shown in the data quality table (see later). Each record with a blank postcode field will generate a warning requiring confirmation from the submitting organisation before the submission can be finalised.
- **Warning** – Data providers are asked to confirm any child postcodes which are the same as the school postcode. In other words, the data provider confirms that the child does live very near to the school and that the school postcode has not been submitted in error.

9. **Duplicate check - Rejection** – The system currently defines duplicates as records within one school with the same: Pupil ref and/or NHS number and/or first name, surname, sex and DOB.

- Individual data entry: it is not possible to save any record defined as a duplicate. Therefore if a user attempts to add a record into a school's pupil grid and enters the same pupil ref and/or NHS number and/or first name, surname, sex and DOB as another record in that school, then the record will be rejected.
- Uploaded data: when uploading records the system checks whether a record is "new" (i.e. not already held) or an "update" (i.e. record already held). The system will check the following fields in order, moving on to the next field if no data is held for the field being checked: NCMP system ID; Pupil Ref (the LA's reference); NHS number; First Name, Last Name, Date of Birth, Sex. Where the system finds a record in the upload file matching one of these fields for a record in the system, and both records have the same URN, the system will update the record in the system. However, the system will not allow any matched records to be uploaded if they create a duplicate record. So if, for example, a record in the upload file has the same NCMP system ID as a record in the system but also has the same Pupil ref and/or NHS number and/or first name, surname, sex and DOB as a third record, then the record will be rejected.

10. **Duplicate measurements check - Warning** – As well as the duplicate check described above, the system will also warn users if they enter exactly the same height and weight combination as the previous record. This may indicate a

² Where:

- 9 can be any single digit number.
- A can be any letter except for Q, V or X.
- B can be any letter except for I, J or Z.
- C can be any letter except for I, L, M, N, O, P, Q, R, V, X, Y or Z.
- D can be any letter except for I, J or Z.
- E can be any of A, B, E, H, M, N, P, R, V, W, X or Y.
- Z can be any letter except for C, I, K, M, O or V.

mistake when manually entering data where the person inputting the data has started to enter the data for another child, but has accidentally duplicated the measurement data for a child which they have already entered into the system.

11. **Same height and weight entered - Rejection** – It is extremely unlikely that a child would have the same height in cm and weight in kg and it is more likely that the height has been duplicated in the weight field by mistake or vice versa. The system will reject all records with an identical height and weight.

Data Quality Table

Data providers also have access to a summary data quality table throughout the collection period. More details on this table are given in the following section as it is also used as part of the post deadline validation.

Post Deadline Validations

This section describes the additional validation and data quality reporting carried out by NHS Digital once all LAs have submitted data and the collection period has ended. These are summarised below and fuller explanations follow:

1. Firstly NHS Digital examines the data quality indicators that the LA has signed off as part of the submission process and queries any which do not match the required conditions.
2. Secondly NHS Digital carries out additional post deadline validations on the record level data.
3. The data quality indicators are published as part of the NHS Digital national report which makes users aware of any LAs who have submitted poor quality data in relation to their peers.

What does NHS Digital do when a data quality issue is identified?

Where issues are identified, NHS Digital contacts the LA concerned and a range of options are available at this time. These include:

- **Allowing the LA to resubmit if there is sufficient time before publication** - This option is only used by exception as once the collection period closes there is usually not enough time for NHS Digital to allow LAs to resubmit data and still publish the national report on time.
- **NHS Digital correcting data after submission** – Again this option is rarely used as outlined in the principles section earlier. An example in previous years was an LA who had submitted all their pupil ethnicity data as one value and realised this was an error when queried by NHS Digital. As there was insufficient time for the LA to collate and resubmit correct ethnicity data for their pupils, a decision was taken for NHS Digital to set all the ethnicity data for that LA to “unknown”.
- **Flagging data as a data quality concern** – this is the most commonly used option. It simply involves adding a data item to the dataset which flags records where there is a concern around the quality of the data. This will have a specific value to alert users as to the reason the record has been flagged and allow them to make an informed decision on whether to include or exclude specific records. For example the 2013/14 data extract had a data quality flag which took the following values:
 - 0 = no data quality issue.
 - 1 = records from one school in London Borough of Enfield which had a high proportion of extreme measurements (60 records).
 - 2 = records from Wakefield Council where one or more measurement warnings were not suppressed at submission (42 records).

Data Quality Table

The NCMP system provides real-time data quality indicators throughout the collection and the LA's NCMP Lead is required to sign off these indicators as being

within acceptable limits as part of finalising their data at the end of the collection. Some of these data quality indicators are presented at LA-level in a data quality table in the annual National Statistics report. This table serves to highlight publicly any LAs which have poor quality data in relation to their peers.

A list of the indicators provided to LAs by the system throughout the collection period is as follows. All the indicators are for year R and year 6 data combined unless specified.

The thresholds used are based on historical analysis of NCMP data and are reassessed for each collection year with a general aim to continually improve the quality of the NCMP data.

A subset of these indicators is published in the national report at LA level and are colour coded green, amber or red depending on the percentage achieved.

1. **Participation rate** – This should be close to 100% and must exceed 85% for both reception year and year 6 separately.
2. **Split between reception and year 6** - The proportion of children entered onto the system in reception year should be between 40% and 60%.
3. **Split by gender** – Within reception year and year 6, the proportion of male children should be between 40% and 60%.
4. **Blank postcodes** – The proportion of blank postcodes should not exceed 5%.
5. **Same pupil and school postcode** - The proportion of children with the same postcode as their school should not exceed 2%.
6. **Extreme measurements** – The proportion of extreme measurements should not exceed 1%.
7. **Whole and half measurements** – The proportion of records where the recorded height is exactly a centimetre or half a centimetre should not exceed 30%. The same check is carried out for weight in terms of a kilogram.
8. **Unknown ethnicity** – The proportion of records with an unknown ethnicity should not exceed 25%.
9. **Same ethnicity** – The proportion of records sharing the same ethnicity should not be 100%. This excludes “unknown” and “not stated”.
10. **Weekend date of measurement** – The proportion of measurements carried out at the weekend should not exceed 2%.
11. **August date of measurement** – The proportion of measurements carried out in August should not exceed 1%.

Additional post-deadline validations

The following additional checks are also carried out. All checks are carried out for year R and year 6 data combined unless specified.

1. **Changes in number measured** – the number of children measured should not be more than 10% different to the previous year. This is carried out separately for children in reception and year 6.
2. **Eligible pupil numbers** - any LA not providing their own headcounts for more than 90% of schools on their list and with a participation rate below 85% will be queried to ensure the system supplied numbers from DfE are correct.
3. **Changes in BMI prevalence** – the prevalence rate for any BMI category should not change by more than 5 percentage points from the previous collection year.
4. **Changes in ethnicity groups** – the proportion of children in each ethnic category will be queried if it was more than 20 percentage points different to the previous year and the ethnic group proportion was in the top five. However, LAs will not be queried if the ethnic categories of “not stated” or “unknown” have decreased by more than 20 percentage points as this represents an improvement in data quality rather than a potential miscoding error.
5. **Schools with a high proportion of extreme measurements** – the proportion of children in a school with an extreme z-score³ should not exceed 10%. This check is carried out separately for height, weight and BMI for both year R and year 6 pupils (but schools with 20 or fewer pupils in that school year will not be queried).
6. **Schools with a high number of extreme pupil postcode to school postcode distance** – the number of pupils in a school where the distance from their home postcode to school is greater than 60km should not be 3 or more.
7. **Schools removed and added by the local authority** – the number of schools removed from their school list by a local authority should not exceed 3 or more than the number added.

³ Measurements are defined as “extreme” when the measurement z-score is lower than -3 or above 4.

Annex A: Calculation of extreme values

Since children's height and weight are dependent on age and sex, height and weight measurements must be standardised to take these factors into account. The standardised value is called a z-score and indicates how far, and in what direction, the measurement deviates from the average (mean) for that age and sex.

High and low z-scores (i.e. measurements that are significantly higher or lower than the mean) are less likely to occur and indicate extreme values.

The NCMP system flags measurements as being 'extreme' if the z-score is less than -3 or more than 4.

To calculate the z-score:

1. Look up child age and sex on the relevant UK90 centiles classification⁴ (there is a separate classification for height, weight and BMI)
2. Retrieve the corresponding *L*, *M*, and *S* values for use in the following formula (where *y* is the measurement e.g. height, weight or BMI):

$$z = \frac{\left(\frac{y}{M}\right)^L - 1}{LS}$$

⁴ There are LMS values for each whole month. Where age falls between two whole months, linear interpolation is used to estimate LMS values. This involves assuming a linear relationship between age and LMS values and deriving the LMS values based on where the age sits between the two months. More information is available at Freeman JV, Cole TJ, Chinn S, Jones PRM, White EM, Preece MA. Cross sectional stature and weight reference curves for the UK, 1990. Archives of Disease in Childhood 1995;73: 17-24 - <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1511167/pdf/archdisch00623-0025.pdf>

Information and technology for better health and care

www.digital.nhs.uk

0300 303 5678

enquiries@nhsdigital.nhs.uk



@nhsdigital

This publication may be requested
in large print or other formats.

**Published by NHS Digital, part of the
Government Statistical Service**

NHS Digital is the trading name of the
Health and Social Care Information Centre.

Copyright © 2016



You may re-use this document/publication (not including logos)
free of charge in any format or medium, under the terms of the Open
Government Licence v3.0.

To view this licence visit

www.nationalarchives.gov.uk/doc/open-government-licence

or write to the Information Policy Team, The National Archives,
Kew, Richmond, Surrey, TW9 4DU;

or email: psi@nationalarchives.gsi.gov.uk