COVID-19 DATA PLATFORM DOCUMENTATION

Data quality assurance

Contents

List of figures

List of tables

About this document

This document has been produced by Ziekenhuis Oost Limburg (ZOL) to fulfil the requirement of the Interreg Euregion Maas-Rijn (EMR) study concerning a COVID-19 Data Platform (CoDaP). The study has three aims:

1. To study the heterogeneity differences in ICU patient populations;
2. To improve clinical decisions guidelines; and
3. To build an IT-infrastructure for secure and efficient data-exchange and analysis.

This document is intended for EMR project reviewers and individuals working on the study on behalf of participating institutions. All comments on this document should be directed to:

- Noëlla Pierlet – Clinical Data Manager at ZOL
- Prof. Dr. Dieter Mesotten – Anaesthetics & Intensive Care at ZOL.

Last updated	24 March 2021
Owners	Noëlla Pierlet, Ben Goethuys and Dieter Mesotten
Author	Joycelyn Harris

Introduction

The goal of the CoDaP study is to set-up a novel, secure IT infrastructure that it is suitable for a multi-hospital, data sharing environment. An aim of the study is to create a visual report that uses COVID-19 ICU patient data. The report should enable participating institutions to study heterogeneity in ICU patient populations and yield insights that that help them prepare for new local outbreaks. To achieve this, the data visualised in reports must be reliable, accurate, consistent and complete. In other words, the data must be of high quality. The purpose of this document is to explain how the quality of CoDaP data is assured. It describes the tools used to evaluate data quality, and identifies the methods employed by the data manager, the data scientist, and the report creator to ensure data quality.

What is data quality

'Data Quality' is a multi-faceted concept. *Data* is facts, concepts, or instructions represented in a manner that is suitable for communication, interpretation, or processing by humans or machine. The International Organization for Standardization (ISO), defines quality as "… the degree to which inherent characteristics of an object meet requirements". Data quality, therefore, relates to how well the data lends itself to communication, interpretation and processing activities.

High quality data can be described as being substantially relevant to the task at hand, timely, accurate, complete and consistent. It can be relied upon to illicit insights that enable an organisation to better coordinate processes. Furthermore, it improves interoperability and communication between stakeholders. Contrastingly, low quality data is inaccurate, incomplete, inconsistent or late, and only passingly relevant to the goals of the organisation. As a consequence, it impedes coordination processes, stifles and misdirects communication between stakeholders, reduces interoperability and undermines organisational goals.

High quality data is paramount to the CoDaP study achieving its goals. The more accurate, complete, consistent and timely that CoDaP data is, the more reliable and relevant the insights that can be gleaned about COVID-19 ICU patients. This in turn leads to decision-making that prioritises and ensures the best use of resources.

Evaluating data quality

Data quality can be evaluated when it is collected, organised, structured and consumed. There are a myriad of methods and tools available that assist with identifying and correcting data errors and flaws. Those adopted in the CoDaP study are: Data quality dimensions, White Rabbit and MS Power BI. However, before the quality of data can be evaluated, one must first identify the data concepts that are under investigation.

CoDaP data concepts

A simply way to identify the data concepts requiring quality assessment is to consider how CoDaP data is produced. Figure 1 contains an overview of the workflow which was first introduced in the document, Sharing data using Microsoft Azure[1]. The workflow begins at point A, with participating

[1] Available from the repository of CoDaP documentation.

institutions providing a Data Extraction File (DEF). The DEF contains data which has been extracted from the institution's electronic patient dossier (EPD) system. The data pertains to ICU patients diagnosed with COVID-19. The data manager for the CoDaP study collates these files for processing by the data scientist (B). Through analysing, reorganising and restructuring the data within the files (C), the data scientist produces a dataset (D). The report creator uses the dataset to build visual reports (E). The reports and the dataset are made available for consumption by participating institutions (F).

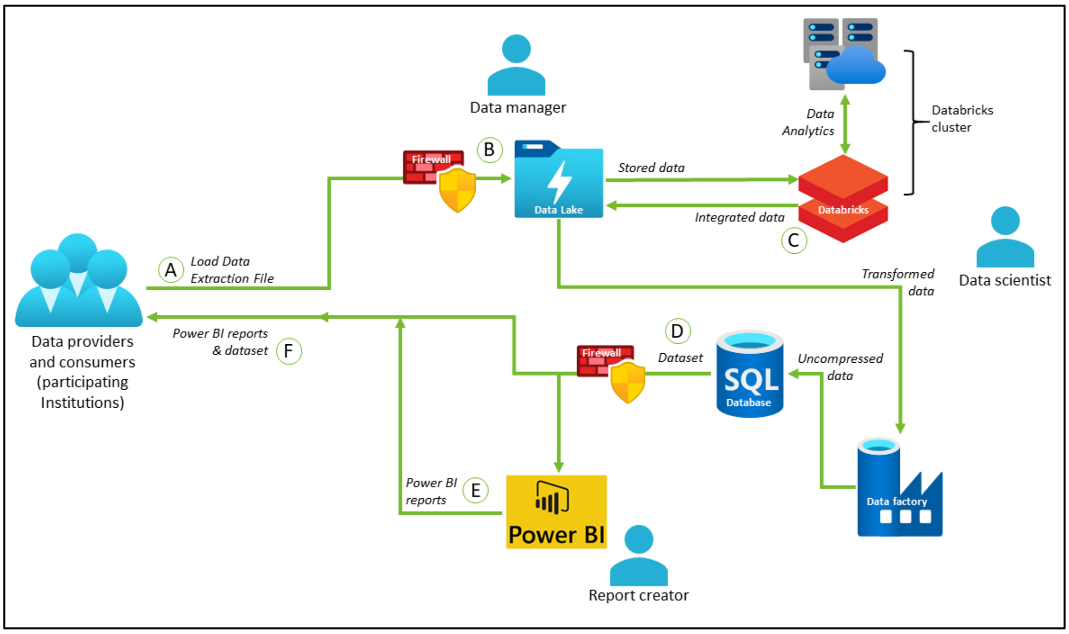

Figure 1 Overview of the CoDaP workflow

The CoDaP workflow produces three types of data artefact: DEF, dataset and visual report. The first two artefacts can be subdivided into data concepts to which quality controls can be applied. These data concepts are illustrated in Figure 2. Every COVID-19 patient that is admitted to the ICU is represented by a *record*. Patients can be distinguished by characteristics such as `Age` and `Weight`. These are *attributes*. Attributes are defined by their *format* which prescribes the datatype, unit of measure and character set. Each occurrence of an attribute is a *data Item*, and the data that it contains is a *data value*. The DEF is a *data file* which is a collection of records stored as a single unit. The CoDaP *dataset* is the aggregation of data files.

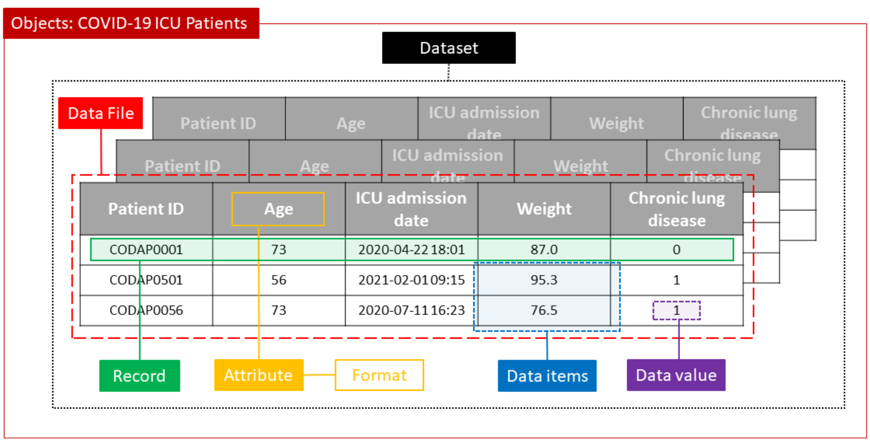

Figure 2 CoDaP data concepts

Data quality dimensions

Every data concept has features that can be measured to indicate data quality. DAMA-NL, a non-profit organisation that develops data management practices, have termed these *data quality dimensions*. One such dimension is *Uniqueness*: the degree to which a real-world object is represented only once. Although there are 60 dimensions available, only a subset of these is relevant for each data concept. For example, a record or a data file can be assessed for its uniqueness, but data values cannot because they are often repeated. The dimensions applicable to CoDaP data concepts are given in Table 1, along with an example that demonstrates noncompliance.

Table 1 Dimensions for evaluating data quality

Data Concept	Dimension	Definition	Example of noncompliance
Objects	Uniqueness	The degree to which real-world objects occur only once as a record in a data file.	A Data Extraction File (DEF) contains 2 records for a patient who was admitted to the ICU only once.
Records	Uniqueness	The degree to which records occur only once in a data file.	A data file contains 3 records for a patient who was admitted to the ICU twice.
	Completeness	The degree to which all required records in the dataset are present.	10 records for patients admitted to the ICU at ZOL have not been included in the DEF.
Attributes	Completeness	The degree to which all required attributes in the dataset are present.	The `Weight` attribute is missing from a data extraction file.
Format	Ability to represent null values	The degree to which a format allows null values in an attribute.	An `ICU discharge date` is given, although the patient is still in the ICU.
	Appropriateness	The degree to which the format is suitable for use.	`Age` is expressed as a decimal number when it should be expressed as a whole number.
Data	Interpretability	The degree to which the language and units of measure of data is appropriate.	`Age` data values are expressed in the number of months instead of the number of years.
	Traceability	The degree to which data lineage is available.	The source of the patient data is unknown.
Data values	Comparability of populations	The degree to which data values representing two populations have the same definition and are measured in the same way.	One DEF expresses `Weight` in kilograms and another in pounds.
	Validity	The degree to which data values comply with rules.	The DEF contains records for ICU patients that were not tested for COVID-19.
	Consistency	The degree to which data values of two sets of attributes within a record or a data file, or between data files, or within a record at different points in time, comply with a rule.	The data value for the attribute, `Lowest P/F ratio`, was recorded within the first 24 hours of admission in one data file and after 48 hours in another.
	Integrity	The degree of absence of data value loss or corruption.	An error message is displayed when the DEF file is opened in MS Excel.
	Metadata compliance	The degree to which the data values are in accordance with their definition, format specification and value domain.	The data value for the attribute, `Gender`, is '3', although the options available are 0 or 1.
	Plausibility	The degree to which data values match knowledge of the real world.	The data value of '300' is given as the `Height` of a patient, which means that the patient was 3 meters tall.
	Precision	The degree to which the error in data values spreads around zero (in statistics).	The standard deviation of `Gender` data values is 10.8. This should be under 0.5 because only two values are allowed.
	Completeness	The degree to which all required data values are present.	A DEF does not contain a `weight` value for each record.
Data files	Completeness	The degree to which all required data files are present.	ZOL failed to produce a DEF.

White Rabbit

White Rabbit is an open source software application for analysing the structure and content of data sources. It is specifically designed to assist data scientists in healthcare environments to prepare files for Extraction, Transformation and Load (ETL) activities. White Rabbit is used in the CoDaP study to evaluate the data in the DEFs and the CoDaP dataset.

White Rabbit uncovers the features of the data under investigation. It scans source data and produces a report containing information about tables, fields and the distribution frequency values. Some of these are depicted in Figure 3, which is an extract from the Field Overview tab of a scan report. The kind of data quality information that can be obtained from a White Rabbit scan report includes:

- The number of rows in a table;
- How many rows are empty;
- A count of the unique values; and
- The standard deviation of field values.

	A	B	C	D	E	F	G	H	I	J
1	Table	Field	Description	Type	Max length	N rows	N rows checked	Fraction empty	N unique values	Fraction unique
2	patients.csv	ID		VARCHAR	36	-1	200	0.0%	200	100.0%
3	patients.csv	BIRTHDATE		DATE	10	-1	200	0.0%	182	91.0%
4	patients.csv	DEATHDATE		DATE	10	-1	200	90.5%	20	10.0%
5	patients.csv	SSN		VARCHAR	11	-1	200	0.0%	200	100.0%
6	patients.csv	DRIVERS		VARCHAR	9	-1	200	16.5%	168	84.0%
7	patients.csv	PASSPORT		VARCHAR	10	-1	200	23.5%	154	77.0%
8	patients.csv	PREFIX		VARCHAR	4	-1	200	21.0%	4	2.0%

Figure 3 Extract from a White Rabbit scan report

The information contained in the scan report enables the user to determine if the data is accurate, consistent, correctly formatted, and so on. For example, it is common for `Gender` to be represented in a data source by the numbers 1 and 2, which related to 'Male' and 'Female'. However, the data source analysed in Figure 4 contains other values for `Gender`. This is apparent because the list is truncated. In other words, there are values present in the `Gender` column that are neither 1 nor 2. With this information, the additional values can be identified and corrected in collaboration with the data provider.

GENDER	Frequency
1	104
2	96
List truncated...	

Figure 4 Truncated list of values in White Rabbit

MS Power BI

The Power Query Editor in MS Power BI enables report creators to prepare data for visual reporting (Figure 5). It analyses columns in a data source to determine the percentage of values that are valid, distinct and unique. It also states the percentage of missing values and values that appear to be erroneous. The report creator can discover further features of the data through viewing the Column Statistics. This provides information such as the minimum and maximum data values, the standard deviation and the value distribution. The report creator's interpretation of these indicators can prompt them to change the data. This is known as *data transformation*. Common data transformation tasks that can be performed via the Power Query Editor include:

- Adding and removing columns;
- Changing data types; and
- Sorting data values.

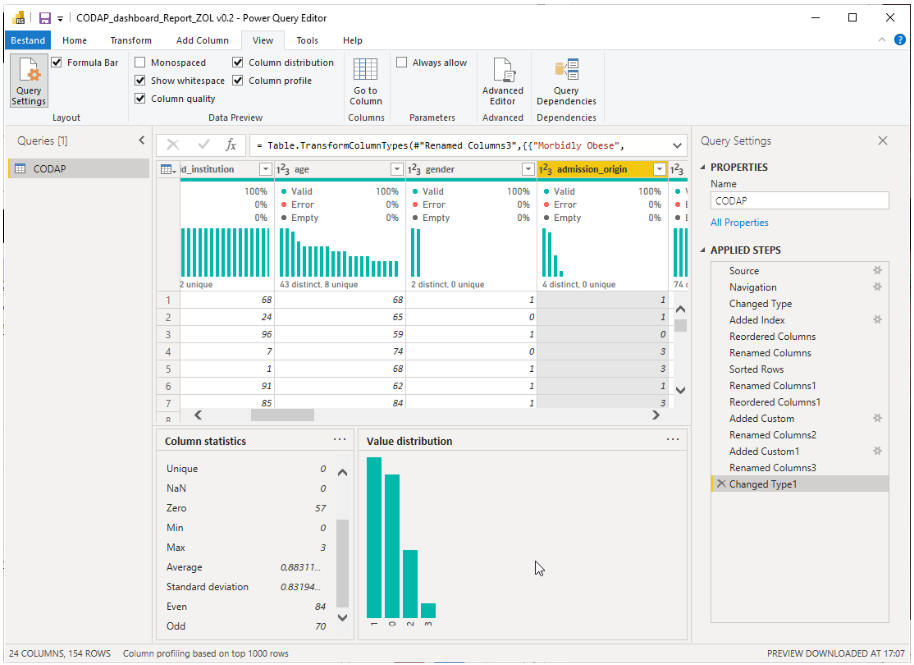

Figure 5 The Power Query Editor within MS Power BI

The transformations applied to the data in MS Power BI do not change the underlying data source. Rather they change how the data is presented in a visual report. For example, the report creator can round down any decimal values given for the attribute `Age` to so that they become whole numbers. These values will be treated as and appear as whole numbers in the visual report, but will remain as decimal numbers in the data source. The Power Query Editor keeps track of the transformations made as *Applied Steps*. The report creator can repeat, re-order and removed these steps. Moreover, the steps persist even after the Power Query Editor is closed or the data source is refreshed.

Ensuring the quality of CoDaP data

The following sections serve to explain how the quality dimensions and tools described in the previous section are used to ensure that the quality of CoDaP data is high. It is stated how the dimension is applied and how quality is measured. Although some dimensions apply to multiple data concepts, there are differences in how it is assessed. For example, both data files and records can be assessed for their *completeness*. However, the former focuses on the availability of required data files and the latter on the availability of required records in a dataset.

Uniqueness of objects and records

The *objects* of the CoDaP study are ICU patients that have been diagnosed with COVID-19. *Uniqueness* in this regard is the degree to which each stay in the ICU is represented only once as a record in a DEF and the dataset. A record that occurs twice is called a *duplicate*. Uniqueness can be evaluated by counting the unique patient identifiers within each DEF and the CoDaP dataset. Data quality tools can uncover if duplicate patient identifiers are present in the data.

Understandably individual data values will be duplicated, such as multiple patients who are 73 years old. It is also anticipated that `Patient ID` values will be duplicated. The `Patient ID` is an identifier assigned to a patient by the supplying institution. Duplications of this value will arise because some patients have stayed more than once in the ICU and re-admitted patients retain their `Patient ID`. Therefore, a low percentage of duplicated `Patient ID` values is to be expected. In these cases, uniqueness can be verified by checking the value for the attribute `ICU admission date`. This date distinguishes a patient's first stay in the ICU from subsequent stays. If duplicate values appear in both the `Patient ID` and the `ICU admission date`, then the data manager and/or data scientist will contact the supplying institution to resolve the duplication.

To further ensure data uniqueness, the attribute `Institution ID` is given to each record on its incorporation into the CoDaP dataset. In the unlikely occurrence that two or more institutions supply records containing equal values for `Patient ID` and `ICU admission date`, then the `Institution ID` will prevent these records from being considered as duplicates. As a final touch, the data manager assigns to each record in the dataset a unique `Index ID`, which serves as a key value in MS Power BI reports.

Completeness

Completeness is the degree to which all required data is present. How completeness is evaluated depends upon the data concept under investigation.

Completeness of data extraction files

When applied to Data Extraction Files (DEFs), *completeness* is the degree to which all required DEFs are present and can be manipulated to create a dataset. Completeness can be measured by counting the number of DEFs available and checking the file format.

To produce a dataset of high quality it is necessary to ensure that all participating institutions provide patient data. Before the CoDaP dataset is complied, the data manager will check that they have received a DEF from each participating institution. There should be 5 DEFs in total. They will also check that the DEF is adequately formatted as a CSV file so that it can be consumed during ETL activities.

Completeness of records

In slight contrast to how it applies to DEFs, *completeness of records* is the degree to which all required records in the dataset are present. This is also commonly known as *coverage* and can be gauged by the number of records that are present in a DEF and in the CoDaP dataset.

The quality management tool, White Rabbit, displays the total number of records present in a DEF file. A figure that is lower than that provided by the supplying institution may indicate that either not all records are present or that the DEF is incorrectly formatted. A figure that is higher than that provided by the supplying institution may indicate that the DEF is incorrectly formatted. The data manager will then take steps to correct the anomaly. This exercise is also applied to the CoDaP dataset. The total number of records it contains should be equal to the sum of records in each DEF file. If this is found to not be the case, then the data scientist will find the source of the incongruency by examining the execution of ETL processes.

Completeness of attributes

In regards to attributes, *completeness* is the degree to which all required attributes are present. Attributes are represented by columns in DEFs and the CoDaP dataset. Data quality tools can calculate the number of columns present in the data. Each DEF file should contain of 59 columns. These are described in the document, Data Dictionary[2]. The CoDaP dataset should contain of 61 columns. Any column count that is higher or lower than these figures indicates that the data is incomplete.

In regards to DEFs, the first step that the data manager will take to resolve incompleteness is to check that the correct separator has been used to distinguish one column from the next. If not, then he or she will correct this. If the correct separator has been used, then the data manager will contact the supplying institution to discuss how data values for the missing attributes can be recovered.

In regards to the CoDaP dataset, missing columns indicate that something has gone awry during ETL activities. The data scientist will then re-examine the execution of ETL processes and correct the irregularity.

Completeness of data values

The *completeness of a data values* refers to the degree to which all required data values are present. Discussions on this topic often refer to *unique* or *missing* values. Unique values are the those that appear only once for a given attribute. A missing value is a value that should be present in the data, but is not. It would be a mistake to assume that all null values are missing values because a null value is often acceptable and sometimes it is even desirable. In contrast, a missing value is always undesirable and it is only acceptable if the data is unobtainable. For example, if the patient survived their stay in the ICU, then a null value is acceptable and desirable for the attribute `Date of ICU Death`. However, if the patient died, then a null value for this attribute is undesirable because it indicates that

[2] Available from the repository for CoDaP documentation.

the patient is still alive. Therefore, it is a missing value. The acceptability of null values is described in the section, Ability to represent null values.

In regards to unique values, completeness can be determined by calculating the number of unique values for a given attribute. The data quality tool, White Rabbit, can assist with this. For each column in the data, White Rabbit shows the number and fraction (expressed as a percentage) of unique values, as illustrated in Figure 6. For some attributes it is expected that the number of unique values will be high. For example, it is expected that over 90% `Patient ID` data values will be unique because very few patients were re-admitted to the ICU. Therefore, an individual `Patient ID` is likely to appear only once in a DEF. In contrast, it is expected the fraction of unique `Gender` data values will be very low, around 2%, because only two values are acceptable: 0 indicating a male patient and 1 indicating a female patient. Therefore, the prospect for unique values appearing in the `Gender` column is low.

Table	Field	N unique values	Fraction unique
CODAP_DataSet_ZOL_20201015.csv	PatientID	97	100,0%
CODAP_DataSet_ZOL_20201015.csv	Age	42	43,3%
CODAP_DataSet_ZOL_20201015.csv	Gender	2	2,1%

Figure 6 Extract from a White Rabbit scan report showing unique values

In regards to missing values, completeness can be determined by calculating the percentages of the possible data values present in the data. For example, the percentage of possible data values for the attribute `Gender` is 100% because it is expected that each patient record will contain the value of 0 or 1 to indicate that the patient was male or female. White Rabbit calculates the fraction of empty values. The higher the percentage, the more values are missing. Figure 7 shows that all values are present for the `Gender` attribute because '0,0%' is displayed in the Fraction empty column. This is not the case for the attribute `BMI` because over 20% of the values are empty. These missing values have arisen because the calculation for `BMI` requires `Height` and `Length` data vales, some of which are also missing from the dataset. The data scientist can seek to correct these and similar concerns in collaboration with the institution that provided the data.

Table	Field	Fraction empty
CODAP_DataSet_ZOL_20201015.csv	BMI	20,6%
CODAP_DataSet_ZOL_20201015.csv	Weight	16,5%
CODAP_DataSet_ZOL_20201015.csv	Lengte	11,3%
CODAP_DataSet_ZOL_20201015.csv	Gender	0,0%

Figure 7 Extract from a White Rabbit scan report showing percentage of missing values

Appropriateness of format

Appropriateness is the degree to which the format is suitable for use. Table 2 shows that the format for the attribute, `ICU admission date`, is the Datetime data type. To ensure that non-technical readers understand what is expected, a simple description of the data type is provided, e.g. *Date & Time*, along with how the value should be expressed, e.g. *yyyy-mm-dd hh:mm* and *2020-04-01 16:15*. The format of each attribute is specified in the document, Data Dictionary[3].

Table 2 Data type and Output format of an attribute

Name	ICU admission date
Simple data type	Date & Time
Unit of measurement	---
Technical data type	Datetime
Output format	yyyy-mm-dd hh:mm
Time of measurement	On ICU admission
Example	2020-04-01 16:15
Mandatory	Yes

Data quality tools display the data type to which an attribute conforms. Furthermore, they signpost values that do not conform to the expected data type. The data manager, data scientist or report creator can correct simple data type errors, such as decimal values given where whole numbers are expected. To resolve more complicated formatting errors, such as the use of text-based values for an attribute which should be expressed in numbers, the data manager and/or data scientist will contact the institution that provided the data.

Ability to represent null values

A *null value* is an empty data item. A zero (0) is not a null value because it conveys information. For example, water freezes at 0˚C. The degree to which a format allows null values can indicate the quality of the data.

A null value is acceptable and sometimes desirable for some attributes because it is not always necessary or possible to take certain measurements. Falling under this category are the attributes `APACHE II`, `Glasgow Coma Scale`, `Ventilation modality`, and also those relating to laboratory results, such as `Adrenaline` and `Dopamine`. Some attributes relating to the date when an activity occurred may also contain a null value. Consider the attributes `Extubation Date` and `Reintubation Date`. A null value for both these attributes indicates that the patient was never intubated. Similarly, the attribute `Discharge Location` may contain a null value if the patient was still in the ICU when the DEF was produced.

A null value is unacceptable for some attributes. For example, it is mandatory to for institutions to provide a `Patient ID` and the `ICU admission date`. A null value in the former impedes Traceability. In the latter a null value indicates that the patient was not admitted to the ICU, in which case their data is irrelevant to the study and should be excluded from the CoDaP dataset.

[3] Available from the repository for CoDaP documentation.

Other attributes for which it would be undesirable if they contained null values are `Admission Origin` and attributes with a BIT (Boolean) format. The `Admission Origin` should not contain a null value because patients never arrive in the ICU without a prior assessment performed by another hospital department. Attributes with a BIT (Boolean) format relate to the presence of comorbidities and treatment therapies, such as `Chronic Liver Disease` and `Vasopressor Use`. Consequently, the data value given should either be 'yes' or 'no'. These values indicate the presence of absence of the comorbidity, or that therapy was or was not administered. However, it must be considered that null values can arise for these attributes if the original value was not entered into the institution's EPD system and cannot be recovered from other data sources within the institution.

Both White Rabbit and MS Power BI display the number of null values that are present for an attribute. Figure 8 shows that there are not any null values for the attribute, `Pulmonary Embolism`. The data scientist can seek to correct undesirable or unacceptable null values through collaboration with the institution that supplied the data.

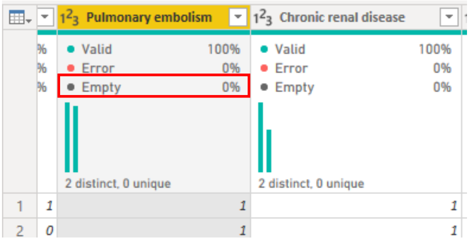

Figure 8 Calculation of null values in MS Power BI

Interpretability

Interpretability is the degree to which data adheres to an appropriate language and agreed units of measure. It was agreed that all data would be presented in the English language in order to ensure that all study collaborators can understand the data. The document, Data Dictionary[4], states the units of measure in which data values should be expressed and gives examples thereof. This is demonstrated by Table 3, which shows the unit of measurement for the attribute, `Weight` and gives an example of how the value should be expressed.

Table 3 Example of unit specification for an attribute

Name	Weight
Simple data type	Number
Unit of measurement	Kilograms (kg)
Technical data type	Double
Output format	---
Time of measurement	On ICU admission
Example	87.0
Mandatory	No

The evaluation of interpretability begins with DEFs. The data manager will check that the agreed units and data types have been applied to the data. They will also correct simple errors, such as the use of the comma punctuation mark as the decimal separator instead of a period mark. The tools and

[4] Available from the repository for CoDaP documentation.

methods they use for this are described in a document produced by Maastricht UMC+. More complex errors are corrected through discussion between the data manager, data scientist and the institution supplying the data. For example, the data manager would be suspicious of `Weight` values exceeding '700' because that the heaviest person in the world weighed 635 kilograms. He or She will suspect that the decimal separator is missing from these values, but they will contact the supplying institution to confirm this before amending the data.

The evaluation of interpretability ends with the review of MS Power BI reports. These reports contain visual representations of the CoDaP data. Clinical and data specialists will review the contents of these reports and alert the report creator to erroneous, unclear or misleading visualisations.

Traceability

Traceability is the degree to which data lineage is available. *Data lineage* is metadata that identifies the source(s) of data and the manners in which data has been organised up to the point of consumption. Data is of low quality when the data lineage the data is unknown.

Traceability is maintained in several ways. Each DEF is stored under the name of the institution that provided it. Every record in a DEF contains the patient identifier that the supplying institution assigned to the patient. If this is not found, then the data manager will discuss this with the institution. In the CoDaP dataset, each record is given an identification number that corresponds to the supplying institution. In MS Power BI, the transformations applied to the data are stored, as illustrated by Figure 5 in the section, MS Power BI.

Comparability of populations

Comparability of population is the degree to which data values representing two populations have the same definition and are measured in the same way. A major aim of the CoDaP study is to compare patient populations from different institutions. It is therefore crucial that data values representing two populations are measured and expressed in the same ways.

The manner in which data values should be expressed and measured is specified in the Data Dictionary[5], as well as the moment when the value should be measured. The range of possible values is restricted for many attributes. For example, values for the `Gender` attribute should be either 0 or 1. This allows a group of males from one institution (i.e. all records where the `Gender` value is 0) to be compared with a group of males from another. Equally, the unit of measure is stated for some attributes so that like can be compared with like. This is the case for the attribute, `Highest FiO2` as shown in Table 4 below. The data dictionary provides when necessary further requirements to which the data value must adhere, such as "*The value should be the highest alveolar-arterial oxygen pressure difference*" and "*The difference is calculated using the formula:* `AaDO2 = 7.13*FIO2 - PaO2 - PaCO2`".

[5] Available from the repository for CoDaP documentation.

Table 4 An attribute where the unit of measurement is specified

Name	**Highest FiO2**
Simple data type	Number
Unit of measurement	Percentage (%)
Technical data type	Int
Output format	---
Time of measurement	Within the first 24 hours in the ICU
Example	80.0
Mandatory	No

White Rabbit can be used to indicate when a data value does not conform to the agreed manner of expression. Such as when the number of unique values for an attribute is shown to exceed the number of expected values. Or if the standard deviation of data values, as illustrate in Figure 9, are suspiciously high or low. The data scientist can seek to correct these and similar concerns in collaboration with the institution that supplied the data.

Table	Field	N unique values	Fraction unique	Average	Standard Deviation
CODAP_DataSet_ZOL_20201015.csv	TotalIntubationDuration_Hours	39	40,2%	297,0365789	145,0984519
CODAP_DataSet_ZOL_20201015.csv	BMI	75	77,3%	30,52584416	4,943497417
CODAP_DataSet_ZOL_20201015.csv	Obesity	3	3,1%	0,47	0,50221655
CODAP_DataSet_ZOL_20201015.csv	Gewicht	41	42,3%	87,70493827	13,24939528
CODAP_DataSet_ZOL_20201015.csv	Lengte	33	34,0%	169,99	10,88062951
CODAP_DataSet_ZOL_20201015.csv	24u_bilirubine	15	15,5%	6081,4	3630,637981
CODAP_DataSet_ZOL_20201015.csv	24u_CRP	92	94,8%	178,1860215	86,65669094

Figure 9 Extract from a White Rabbit scan report showing standard deviation of data values

Consistency of data values

Consistency is about comparing two or more data values and determining how consistent they are with each other. This is often referred to as *soundness*. A general example of inconsistent data is a record that states that a company's city location is 'Paris' and that the country location is 'Belgium'.

Consistency is very important for data values that depend upon or are derived from other values. For example, the `BMI` value in a record must be consistent with the `Height` and `Weight` values because its calculation is contingent upon these values. Or for patients that received `Invasive ventilation`, the record should not contain a null value for the `Intubation Date` because the former took place on a particular day. Similarly, when a value is provided for `Date of ICU Death`, a value should be provided for `Date of ICU Discharge` and also for `Discharge location` because the patient will have been directed to a particular place.

The consistency of attributes in DEFs is evaluated by the data manager. The tools and methods used for this are described in a separate document produced by Maastricht UMC+. They will seek to correct inconsistencies in collaboration with the institution that provided the data.

Integrity of data values

Integrity is the degree to which the absence of data values signifies data loss or corruption. This can be assessed by determining if a data file or dataset contains the expected number of records and attributes. Much like Completeness of attributes described previously, White Rabbit and MS Power BI both count the number of records (rows) and attributes (columns) present in DEFs and the dataset. Any count that is higher or lower than expected can indicate that the integrity of the data is compromised.

In regards to DEFs, the first step that the data manager will take to ensure data integrity is to open the file. If the DEF is unopenable, then he or she will contact the providing institution to obtain a new DEF. The next step is to check that the correct separator has been used to distinguish one attribute from the next. If not, then the data manager can correct this. If the correct separator has been used, then the data manager will contact the providing institution to obtain data values for the missing attributes.

In regards to the CoDaP dataset, missing columns indicate that something has gone awry during ETL activities. The data scientist will then re-examine the execution of ETL processes and correct the irregularity.

Metadata compliance

Metadata compliance refers to the degree to which the data values are in accordance with their definition, format specification and value domain. The manner in which data values should be expressed are specified in the document, Data Dictionary[6]. This includes the unit of measurement, data type and minimum and maximum values. Data quality tools can be used to identify noncomplying data values as described in the section, Comparability of populations. The data scientist can seek to correct these issues in collaboration with the institution that provided the data.

Precision of data values

Precision in regards to data values refers to statistical notion of distribution. It is the degree to which values spread or disperse around zero. White Rabbit can be used to uncover this. It can perform statistical calculations for all integer, real and date data types, and display the values in a scan report. The calculations available are:

- Average value;
- Standard deviation (sampled);
- Minimum and maximum value; and
- Quartiles (sampled) at 25%, 50% (median) and 75%.

The data manager defines in the White Rabbit scan settings menu the number of values that will be stored for the calculation of numeric statistics. This is the *reservoir size* and it is shown in Figure 10. These values are randomly sampled from the data values in the scan report. If the number of values is smaller than the reservoir size, then the standard deviation and three quartile boundaries are the exact population statistics. Otherwise, the statistics are approximated based on a representative sample.

[6] Available from the repository for CoDaP documentation.

The average, minimum and maximum are always true population statistics. For dates, the standard deviation of dates is given in days. The other date statistics are converted to a date representation. When any of these statistics appear too high or too low, the data scientist can seek to clarify the reasons behind this with the institution that provided the data and, where possible, take steps to correct it.

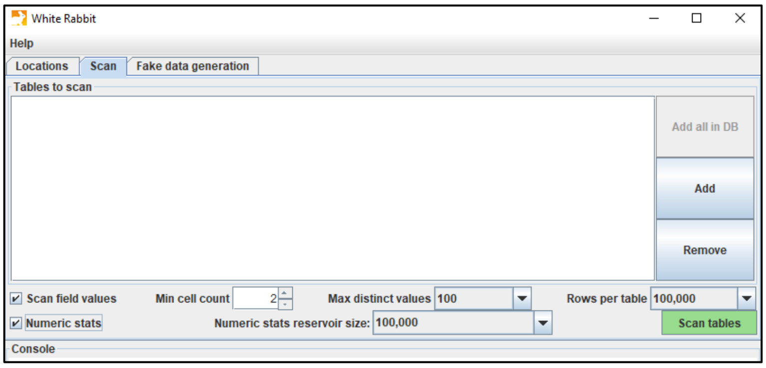

Figure 10 White Rabbit scan settings

Validity of data values

Validity refers to the degree to which data values comply with the rules by which it is defined. A valid data value is one which complies with the permissible values of the attribute.

It is important to note that validity does not indicate accuracy. For example, the provision of the value '42' for the `Age` attribute is valid because it is a whole number that falls within the range of a typical human lifespan. However, if the patient for which the value was recorded is actually 52 years old, then the value is valid but inaccurate.

When evaluating data validity, one must not only consider completeness (i.e. the extent to which all required data values are present), but also consistency amongst values, as well as values' adherence to the attribute definition, format specification and value domain. Data validity can be determined by calculating the percentage of data values that are incomplete, inconsistent and do not comply with attribute descriptions. As such, the methods described in the sections Completeness, Appropriateness of format and Consistency of data values are also used to determine data validity.

Plausibility of data values

Plausibility refers to the degree to which data values correspond to knowledge of the real world. The first party that evaluates this is the institution supplying the data because its employees have directly interacted with the real-world object, i.e. the patient.

Value plausibility can be further evaluated is by comparing it to the data limits for the attribute. White Rabbit can be used to calculate the minimum, average and maximum values available in the data for a given attribute, as demonstrated in Figure 11. Combining this information with previously accrued knowledge, the data scientist can reasonably infer the plausibility of a value. For example, if the median average `Height` of patients is 170 (centimetres), then a `Height` data value over 250 will cause the data scientist to question if the patient was indeed that tall. When a value appears to exceed the data limits, the data scientist can seek to uncover the reason with the institution that supplied the data.

Table	Field	Average	Min	Median	Max
CODAP_DataSet_ZOL_20201015.csv	Height	169,99	131	170,5	193
CODAP_DataSet_ZOL_20201015.csv	Age	66,86	34	70	88

Figure 12 Extract from a White Rabbit scan report showing data limits

The presentation of data in a MS Power BI report affords individuals within participating institutions the opportunity to identify implausible data. At a basic level, it is likely that report viewers will detect implausible outliers or data groupings that lie outside of their expectations. In such cases, the viewer should raise their concern with the data scientist and/or the report creator. Additionally, the process for obtaining feedback from report viewers can be formalised by defining user stories. These should contain questions that viewers should be able to answer by interacting with the report. For example: *How many male patients were morbidly obese?* If viewers are unable to answer a question, then the report creator must find an alternative approach for displaying the data and revise the report accordingly.

Glossary

Attribute

A characteristic or distinguishing feature of a real-world object.

Business Intelligence

Comprises of strategies and technologies used by an organisation for the data analysis of business information.

CoDaP

An acronym for *COVID Data Platform*.

Concept

Unit of knowledge created by a unique combination of characteristics.

CSV

An initialism of Comma *Separated Values*. It refers to a plain text file and is normally used to list tabular data, where each line is a single data record and datapoints (i.e. values) are separated by a single character.

DAMA-NL

An initialisation of The Data Management Association International Netherlands; The Dutch branch of an international not-for-profit organization of data management professionals which focuses on the development and execution of architectures, regulations, practices and procedures that support the data management lifecycle. For further information visit http://www.dama-nl.org/data_quality/

Data

A representation of facts, concepts, or instructions in a formal manner, suitable for communication, interpretation, or processing by humans or by automatic means. It is a collective term for several concepts such as records, attributes, data values and metadata.

Data concept

A form by which data is structured and organised in an information system. 'Dataset', 'record' and 'attribute' are examples of data concepts.

Data file

Data stored on a computer as one unit with one name.

Data Item

One occurrence of an attribute.

Data lineage

The metadata that identifies the sources of data and the transformations through which it has passed up to the point of consumption.

Data Model

An abstract model that organizes elements of data and standardizes how they relate to one another and to the properties of real-world entities.

Data value

The value of a data item that forms part of a record.

Dataset

A data structure normally presented in a tabular pattern, where each column describes a particular variable and each row corresponds to a given member of the dataset. It points to or references the data stored elsewhere.

DEF

An abbreviation for Data Extraction File. It is the CSV file that contains datapoints (values) pertaining to patients infected with COVID-19 that have been admitted to an ICU.

Dimension

In data quality, a 'dimension' is a measurable characteristic. In business intelligence, it refers to a category for summarizing or viewing data.

ETL

An initialism of Extract, Transform, Load which is a general procedure for copying data from one or more sources into a destination system which represents the data differently from the source(s) or in a different context than the source(s).

Format

A form by which data is structured and organised in an information system that is a part of the specification of an attribute.

Institution

A hospital that is participating in this CoDaP project.

ISO 9000

Sets out the criteria for a quality management system based on a number of quality management principles, including a strong customer focus, the motivation and implication of top management, the process approach and continual improvement.

MDS

An initialism of *Minimal DataSet, which* is the core record of the CoDaP study and consists of fields for that structure data related to demography, admission and discharge, physiology, and primary and secondary outcomes.

MS

An initialism of Microsoft.

MUMC+

An initialism of Maastricht University Medical Centre+, a healthcare provider that is a participating in the CoDaP project.

Object

Anything perceivable or conceivable. An object is characterised by Properties and represented by records.

Participant

An individual contributes to the CODAP project in a professional capacity.

Power BI

A suite of business intelligence, reporting, and data visualization products and services for individuals and teams developed by Microsoft.

Property

A feature of an object that has real value and is recorded by a data value.

Quality

The degree to which inherent characteristics of an object meet requirements (ISO 9001).

Record

A logically related set of values that represent a (real-world) object which states results achieved or provides evidence of activities performed. It forms part of a data file and is composed of data values.

Reference data

Data used to categorize other data. An example is the CoDaP Data Dictionary.

Report

In Power BI, a *Report* is a multi-perspective view into a dataset, with visuals that represent different findings and insights from that dataset. A report can have a single visual or pages full of visuals.

Requirement

A need or expectation that is stated, generally implied or obligatory (ISO 9000).

Standard deviation

The average amount of variability in your data set. A high standard deviation in normal distributions indicates that data values are far from the mean, while a low standard deviation shows that values are clustered close to the mean.

Table

An arrangement of data in rows and columns.

Validation

Confirmation, through the provision of objective evidence, that the requirements for a specific intended use or application have been fulfilled.

Value domain

A set of permissible values of an attribute.

Verification

Confirmation, through the provision of objective evidence, that specified requirements have been fulfilled.

Visual

See **Visualization**.

Visualization

Also known as a *visual* this is a graphical report element in Power BI that displays insights discovered in the data. A Power BI report might have a single page with one visual or it might have pages full of visuals.

White Rabbit

An open source quality management application that can be used to analyse the structure and contents of a database. For further information visit https://github.com/OHDSI/WhiteRabbit.