INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Data Quality Assessment in Healthcare Data: A Case Study

Stephanie Cardoso Monteiro

Master's in **Integrated Business Intelligence Systems**

Supervisor:
PhD João Carlos Amaro Ferreira, Assistant Professor with Habilitation,
ISCTE-IUL - Instituto Universitário de Lisboa

Co-supervisor:
PhD Bruno Moisés Teixeira Oliveira, Adjunct Professor,
ESTG-PP - Escola Superior de Tecnologia e Gestão - Politécnico do Porto

December, 2023

TECNOLOGIAS
E ARQUITETURA

Department of Information Science and Technology

Data Quality Assessment in Healthcare Data: A Case Study

Stephanie Cardoso Monteiro

Master's in **Integrated Business Intelligence Systems**

Supervisor:
PhD João Carlos Amaro Ferreira, Assistant Professor with Habilitation,
ISCTE-IUL - Instituto Universitário de Lisboa

Co-supervisor:
PhD Bruno Moisés Teixeira Oliveira, Adjunct Professor,
ESTG-PP - Escola Superior de Tecnologia e Gestão - Politécnico do Porto

December, 2023

To my dear father and mother,
this work is the fruit of our struggle and our victory.
To those who taught me the value of knowledge

Acknowledgment

I begin this thesis by expressing my deepest gratitude to everyone who, in one way or another, contributed to the realisation of this work. Helen Keller has said, "Alone we can do so little; together we can do so much".

I start by expressing my deepest gratitude to my beloved father, Mario Monteiro and my beloved mother, Maria Alice Monteiro, those who have walked this path with me since day one, giving me love and strength in a way no one else could give. I can not put my gratitude into words.

To my supervisor, professor Bruno Oliveira and professor João Carlos Ferreira, for all the availability, guidance, collaboration, and knowledge shared with me during this process, my sincere gratitude.

To my professor, Luís Elvas, who also accompanied me during the process, being always available for help and guidance, even though he was not part of the official team, thank you very much.

To Dr. Luís Rosário, I appreciate all the availability being the bridge between ISCTE and the hospital during my research. This work would not be possible without his collaboration. To my dearest love, Massochy Ventura, who countless times took me in at times of anguish, lifted me up when I had no strength left, and spent late nights by my side, I can't thank you enough. Walking with you is kinder, gentler and more hopeful. Thank you for holding my hand and sharing my dreams.

To my dearest sisters Ivanusa and Ester, with whom I share countless moments of joy, having you as a sister is a blessing. I feel your support, and I appreciate it. To my dearest little brothers, Samuel and Samy, who have encouraged me all the way long. Having you is the best present my parents could give me.

To my beloved oldest bother, Admir, the one who has been my idol since my first degree. The admiration and respect I have for you I could never express in words. Thank you for your unconditional support.

To friends and colleagues, I express my gratitude for all the support, comprehension, and motivation. This adventure would not be the same without each of you.

Resumo

Dados com qualidade são essenciais para monitorizar e avaliar as atividades do negócio. Quando se trata de domínios críticos como a área da saúde, a qualidade dos dados tem um impacto fundamental na prestação de serviços mais precisos e rápidos. Considerando cenários epidemiológicos como a COVID-19, os dados assumem um papel essencial no apoio às respostas sociais dadas pelos decisores. Para tal, a partilha e a visão integrada dos dados permitem identificar as melhores abordagens e os sinais críticos que podem conduzir a melhores diagnósticos e tratamentos, entretanto, lidar com a extração e integração de dados provenientes de várias fontes não é uma tarefa fácil e implicam inúmeros desafios relacionados com a acessibilidade, representação e interpretação dos mesmos. Diferentes problemas relacionados à qualidade dos dados podem ser levantados e quando não tratados corretamente, podem pôr em causa à tomada de decisões.

O principal contributo desta tese é avaliar a qualidade de um conjunto de dados de um hospital português aquando utilizados para integração com repositório de dados partilhado no âmbito de um projeto europeu. Estes dados foram analisados, identificando as principais características e problemas. Os problemas identificados foram posteriormente mapeados com a respetiva dimensão da qualidade de dados violado. Regras foram definidas servindo como diretrizes para auxiliar na correção dos problemas e prevenir que os mesmos ocorram futuramente. Para efetuar esta avaliação, propôs-se uma metodologia que avalia a qualidade dos dados a dois níveis, a nível das regras e das dimensões da qualidade dos dados, calculando posteriormente o *score* relativamente a qualidade dos dados avaliados. Os resultados foram discutidos e avaliados.

Keywords: Dados de saúde, Qualidade de dados, Integração de dados e Avaliação da qualidade dos dados

Abstract

Reliable data is essential for monitoring and evaluating business activities. When critical domains such as Healthcare are involved, data quality has a crucial impact on delivering more accurate and fast healthcare services. Considering epidemiological scenarios such as the COVID-19 pandemic, data can assume an essential role in supporting social answers carried on by the primary decision-makers. For that, sharing and having an integrated view of the data allow for identifying the best approaches and critical signals that could lead to better treatments and diagnoses. Nevertheless, leading with data extraction from several sources is not an easy task and can lead to enormous challenges related to data accessibility, representation, and interpretation. Several data quality problems can occur and, when not adequately addressed, can question the decision-making support.

The contribution of this thesis was to perform a data quality assessment from a subset of data from a Portuguese hospital when used in the context of integration to a common shared repository within the scope of a European project. A deep data profiling analysis in the source database was conducted, identifying the main characteristics and, afterwards, the main issues. Each issue was later mapped with its corresponding data quality violation, and rules were defined as guidelines to address these issues and prevent future ones. To classify the quality of the source data, a methodology was proposed to evaluate the data into two levels, quality roles level and data quality dimensions level, calculating the data quality score. The final results are discussed and evaluated in this work.

Keywords: Healthcare Data, Data Quality, Data Integration and Data Quality Assessment

Contents

List of Figures

List of Tables

List of abbreviations

DCC: Data Coordinator Center.
DQ: Data Quality.
ECG: Electrocardiogram.
EHR: Electronic Health Record.
ETL: Extract Transform Load.
FK: Foreign Key.
GE: Great Expectation.
HES: Hospital Episodes Statistics.
HTML: HyperText Markup Language.
JSON: JavaScript Object Notation.
NCI: National Cancer Institution.
NHS: National Health Service.
NLP: Natural Language Processing.
PK: Primary Key.
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses.
QN: Question Number.
RL: Quality Rule.
RWD: Real Word Data.
SQL: Structured Query Language.

CHAPTER 1

Introduction

A vast quantity of Real Word Data (RWD) is provided by health care services, and they represent around 30% of the data that is kept worldwide [**2**]. This amount of data may be due to the implementation of Electronic Health Record (EHR) all over the world to support the management of patient's records and keep track of their health [**3**]. With the massive expansion in health technology, thousands of bytes of information are stored daily in EHR systems, subjecting the healthcare area to new challenges and developments [**4**].

The healthcare data plays an essential role in the entire life cycle of a patient's treatment, from the moment a patient enters a facility to the diagnosis and dismissal, including current treatments, health history, medication allergies, insurance information, and so on [**5**]. This increase in data opened opportunities for data-driven decisions and highlights the importance of data quality, which is essential in healthcare since the wrong or poor data quality puts lives at risk. For instance, "duplicated drug orders entered into two separate prescribing systems used by a hospital resulted in nurses administering an excessive amount of insulin to a patient, resulting in death" [**6**].

Despite the primary goal of the EHR systems being the management of the patient's health information, the amount of data storage boosted a paradigm-changing, calling attention to the secondary usage of this data in the healthcare research field [**7**]. When we talk about the research field in healthcare, the COVID-19 pandemic reaffirms how much value the use of EHR data can add. However, that can bring significant challenges, essentially the potential for poor data quality [**8**]. Research has highlighted that data collected in the context of patient care might be of inferior data quality than data collected in scientific studies [**7**], which may imply an extra effort to analyse and determine whether such data can be used in each research.

1.1. Motivation and context

COVID-19 is a global pandemic that has caused millions of deaths worldwide. When such a disaster strikes the world, individual interests are often put aside, and the whole world starts working to solve such a problem. Countries collaborate on numerous initiatives and projects to bring meaningful insights that can add value to finding cures and understanding the consequences of such diseases, establishing patterns and correlations between other diseases. One of these global initiatives is the "Capacity Project" [**9**], which aims to collect data regarding cardiovascular complications in patients diagnosed with COVID-19

to study their possible correlation. The collaboration of different facilities would provide more data to the project, allowing a more in-depth study.

Recognising this initiative's importance and the benefits it would bring to the medical field, the Hospital of Santa Maria decided to also contribute to the project by providing the necessary data. However, ensuring their quality is crucial for this data to be used.

Having said that, this academic work is motivated by the need to evaluate the quality of the available data to define the necessary steps to guarantee the required quality, thus ensuring that the data reaches its destination, fulfilling the requirements.

1.2. Objectives

Considering that the object of the "Capacity Project" is to study the relationship between patients with cardiovascular diseases and COVID-19, this academic work has the following objectives:

(1) Identify the patients with cardiovascular complications diagnosed with COVID-19 and determine whether they are eligible for the Capacity study.
(2) Do the data profiling in the main entities used to identify such patients, highlighting the main issues of each one of them.
(3) Define the rules to guarantee the data quality requirements of each entity.
(4) Evaluate the data quality of the source database based on analytical questions representing important data quality requirements that can compromise its usage.
(5) Calculate the data quality score.

1.3. Dissertation structure

With the motivation and the objectives of this thesis presented, a brief presentation of the remaining chapters is conducted in this section. This academic work is composed of 5 chapters (including the Introduction):

Chapter 2 - Literature review: This chapter examines the existing literature and research within the field, encompassing the investigation methodologies, search strategies and inclusion criteria, search result and document selection and the related work

Chapter 3 - Santa Maria Hospital Case Study: The primary step to understanding and analysing the Santa Maria Hospital database occur in this chapter. A data profile analysis is performed, indicating the main issues encountered and mapping them to their corresponding data quality issue. Also, a set of quality rules to ensure the quality of these data is proposed.

Chapter 4 - Data Quality Assessment: In this chapter, a data quality assessment occurs based on five proposed analytical questions. Each question's final data quality score is calculated, and the outcome is analysed.

Chapter 5 - Conclusion: The final chapter of this academic work summarises the key findings and provides a comprehensive discussion of the results. Additionally, directions for future work are addressed.

CHAPTER 2

Literature Review

This chapter examines the current landscape regarding data quality in healthcare. The main focus is exploring research that addresses the leading data quality problem faced when integrating data in healthcare. Yet a comprehensive analysis of the existing literature related to the proposed methodologies and frameworks to evaluate the data quality is presented based on Preferred Reporting Items for Systematic Reviews and Meta-Analyses methodology,PRISMA.

2.1. Investigation methodology

To understand the relevance of this study, it was crucial to investigate the existing literature in the field of DQ in healthcare. In this chapter, the main result of the investigation is addressed, starting by explaining the methodology used for conducting such research and then presenting the main insight obtained through the process. Analysing Figure 2.1, we can notice from the trend line that there is growth on the topic in the study, reinforcing its relevance.

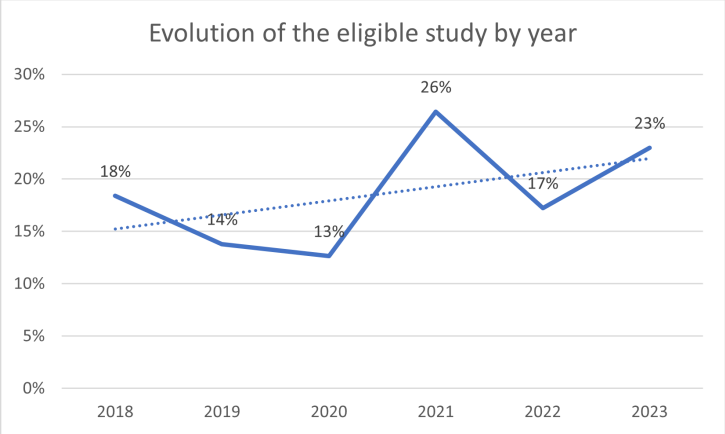


FIGURE 2.1. Evolution of the eligible studies by year

This systematic literature review followed the PRISMA[1] (Preferred Reporting Items for Systematic Reviews and Meta-Analysis) Methodology. The search was performed using two repositories, namely *Scopus*[2] and *Web of Science*[3] with the search question *"What is the state of the art in Data Quality in the Healthcare field?"*

[1]http://prisma-statement.org/
[2]https://www.scopus.com/
[3]https://www.webofscience.com/

2.1.1. Search strategy and inclusion criteria

As previously mentioned, the search was conducted using two databases, Scopus and Web of Science, from April to June 2023. It only considered articles, conference papers and reviews published between 2018-2023 written in English. The document collected was only related to the computer science and engineering field. The search strategy was based on one query, one associated with DQ in the healthcare field, considering the concept("Data Quality"), the context ("Healthcare") and the population ("Evaluation" or "Assessment").

2.1.2. Search result and document selection

Having applied the search strategy defined in 2.1.1, 87 documents from *Scopus* and 60 documents from *Web of Science* were obtained, summarising 147 documents, including the duplicate ones. These documents were exported to Zotero[4], where the 30 duplicate was eliminated, remaining only 117 documents. The first analysis of these documents was reading only the title and the abstract, resulting in the disposal of 62 documents since they didn't fit the scope of this work. The remaining 55 documents were read, and 19 were included in this literature review. Figure 2.2 synthesises the selection process of the total article studied.

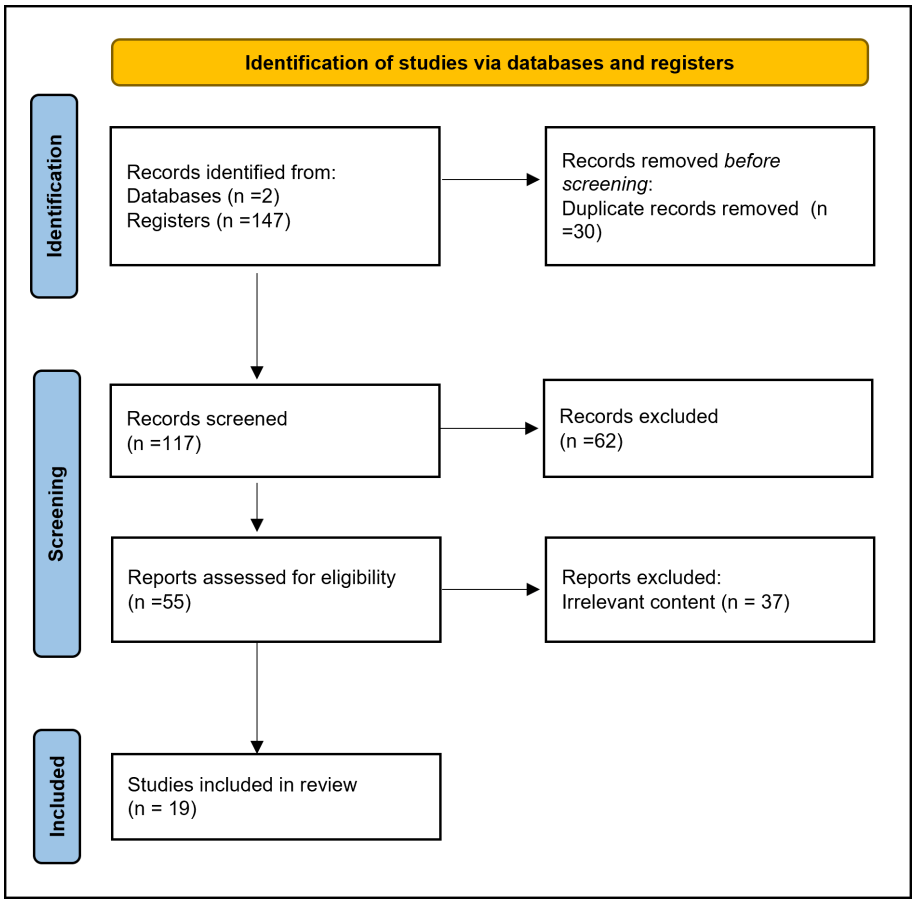


FIGURE 2.2. PRISMA workflow diagram. Fonte: [**1**]

[4]https://www.zotero.org/

2.2. Related work

The EHR system has yielded unquestionable benefits to medical facilities in healthcare by diminishing administrative tasks and facilitating data accessibility. Massive amounts of data about patients, symptoms and diseases or laboratory results are collected daily and stored in different formats and systems [**10**]. Although this system has brought many advantages, the amount of data generated can extend beyond assisting the organisation's daily tasks. Despite the primary goal of the EHR systems being the management of the patient's health information, the amount of data storage boosted a paradigm-changing, calling attention to the secondary usage of this data, such as for research purposes and essentially supporting decision-making [**11**].

When mentioning the research field in healthcare, the COVID-19 pandemic reaffirms how much value inter-organizational collaboration can add and the importance of data-sharing. The pandemic has made it increasingly evident that data-driven improvements in healthcare are crucial for enhancing service quality and obtaining responses from health-care systems that contribute to saving more lives [**12**]. In healthcare, research is often motivated by the common interests of different organisations, countries or even the overall population. For that, collaboration is required, such as sharing data between the involved parties. Since different organisations implement different health information systems to support the management of patient records and keep track of their health [**13**], the data usage in the research often needs to be integrated into a common repository where the research will occur.

According to [**14**], the process of integrating data from multiple sources into a single repository can be commonly known as the ETL process (Extraction, Transformation and Loading), and it's considered one of the most relevant components when talking about populating a repository with a large amount of data such as Data Warehouse and represent a significant effort in data integration projects. In the ETL process, the transformation phase, which includes data cleansing, is considered one of the most challenging since data comes from heterogeneous systems with different formats and meanings. An estimated 40% of these data is compromised in some manner [**15**]. A study highlights that when referring to the development of analytical systems, in which the ETL phase is seen as a crucial one, the data quality subject is taken as the primary obstacle to the success of such project and the responsibility for guaranteeing the quality of the integrated data lies on the ETL process [**16**].

Different data quality problems are usually identified in the ETL, for instance, lack of integrity constraints resulting in violation of the primary keys, embedded values when multiple values are entered into a single attribute, duplicate records, missing values, variety of data types, naming conflict when we have different data sources with synonymous attributes, syntax inconsistency and so on [**16**]. This problem can question the quality of the integrated data when not adequately addressed and subsequently cause direct or indirect impacts on the business decisions that rely on it [**16**].

In conventional use cases, the source data is either structured (typically in relational databases) or semi-structured (such as in spreadsheets), making it relatively easy to identify relevant information. Nevertheless, it is common to encounter unstructured data in this so-called "big data era" and the advancement of artificial intelligence algorithms for data processing. This is particularly evident in technical fields that deal with complex corpus. For instance, the healthcare sector is marked by highly specialised terminology and textual codification of fundamental topics, requiring a certain level of literacy to comprehend and scrutinise the context and data. A widespread issue in healthcare data is ambiguity since it's common to have different ways of describing diseases, diagnoses and symptom [**17**].

According to [**18**], a project was conducted for the National Cancer Institution (NCI) in which the main goal was to set up a Data Coordinator Center (DCC) to develop a centralised data repository for different research purposes. This study emphasised data conciliation as one of the critical steps for the project's success but also the main challenge since data are extracted from different sources. Even when discussing similar data, the code and storage methods can differ, including various formats. This fact can be a challenge when understanding how data are coded and if missing or unavailable data is a question.

Another concern when integrating healthcare data is that it usually includes medical images and free-text reports, which generate a large amount of unstructured data. This fact usually brings significant challenges to extracting consistent and meaningful information [**17**]. Creating structured representation from non-structure data to make it more understandable and usable for knowledge acquirement can require substantial effort [**19**] and the usage of Natural Language Processing (NLP) is frequently adopted as a solution to this issue [**17**], [**20**].

2.2.1. Data quality dimensions

Data quality (DQ) measurement has been essential to ensure confidence in data-driven decisions. When we mention the healthcare field, where the data-driven decision can represent the difference between the life and death of several patients, the importance of reliable data is even more crucial [**21**]. Since the 80s, research about DQ has been conducted and is no "longer a question of hygiene", but a necessary process to ensure operational excellence and it is often associated with the "fitness for use" principles, which refers to the subjectivity and context-dependency of the topic [**22**].

The DQ concept is mentioned as multi-dimensional, where each dimension describes a single aspect of the quality of the data, and it can be measured using different DQ metrics. When concerned with the DQ dimension, a wide variety of dimensions has been proposed over the years. However, despite all research and ongoing discussion related to the topic, there is still no consensus on which dimensions are the main ones for DQ measurement, and that can vary according to the area or subject in concern [**22**]. According to the [**23**], [**24**], they can be included in four main categories: *intrinsic* - explaining data having

quality in their own right, *accessibility* - highlighting features related to the access of the data, *contextual* - defending that quality is evaluated depending on the task on hand; and *representational* - valuing aspect associated with the meaning of data and its format. Different studies have taken place in the healthcare context, and some of them indicate the DQ dimension as the parameter for assessing the quality of the data under analysis.

According to [**23**], an investigation was made regarding data governance in healthcare, highlighting the impact of DQ dimensions in the data governance process. Various DQ dimensions were identified and explored, understanding how each contributes to the overall data quality. This study focuses on understanding which of these dimensions makes the most sense in the context of big data in healthcare. The in-depth research has led to the understanding that accurate and reliable data is essential for informed decision-making and health research. Poor quality data can lead to erroneous conclusions, compromised patient care and ineffective public health initiatives.

In England, a study was conducted using the Hospital Episodes Statistics (HES) dataset as the key data source [**25**], which contains data for all hospital admissions of NHS patients. This exploratory study aimed to provide a novel assessment and analysis of the data quality of this database "in the recording of mandatory diagnoses for patients with autism, type II diabetes mellitus with peripheral complications and Parkinson's disease dementia". To study the consistency of this data set, a machine learning algorithm was used, identifying important predictors associated with data inconsistencies using a random forest classifier. A full report of the data inconsistency was made available, highlighting the causes of these mistakes.

In the context of COVID-19, a comprehensive data quality model was designed to assess public COVID-19 big data sets in Economic and Monetary Community of Central Africa region countries. To define the main characteristics of the data, a data quality model was proposed, taking as a basis the four categories of data quality, namely intrinsic, accessibility, contextual and representational. A framework was proposed, and one of the main modules of the framework consists of a data profiling task to evaluate what they call the four adequacies (4A): contextual, temporal, operational and explanatory adequacy. To determine the level of adequacy in each category, some data quality dimensions such as accuracy, consistency and completeness were considered using proper metric to evaluate them. The model encompasses key dimensions of data quality and provides a systematic approach for evaluating and improving data quality in that context, highlighting the importance of data quality in informing effective decision-making and offering insights into the practical application of the model through a case study [**26**].

When referring to DQ assessment, the DQ dimension is an essential characteristic to validate. They are used to measure, quantify and manage data quality [**27**]. According to [**6**], despite the lack of consensus on the best dimensions to consider when analysing DQ, they propose seven dimensions as the main ones to evaluate DQ in the healthcare context. The first mentioned DQ dimension is *Accuracy*, which "measures whether data

was recorded correctly and reflects realistic values", [**27**]. To ensure that, it's essential to ensure that the data represents the correct data of the intended attribute. For instance, suppose a patient's actual weight is 80 kilograms, but due to a typographical error or misreading, the weight is recorded as 800 kilograms. This data is inaccurate and does not reflect realistic value. [**27**] yet mention that the level of accuracy of a given data set is calculated by dividing the number of fields judged "correct" by the number of the total fields tested.

Another DQ dimension pointed out is *Completeness*; data must include all the relevant details [**28**] and can be analysed using missing values of that attribute [**27**]. In healthcare, completeness can be analysed as coverage of baseline features or data required for a particular disease, patient or treatment, for example, or referring to the right amount of available data. For instance, a patient's name must include the first and last name, at least [**6**]. The presence of null values usually identifies the completeness of a value, [**27**], and a specific metric is used to calculate it, such as the "number of not null values" divided by the "total number of values". However, [**21**] calls attention to the meaning of the null values since they can have different origins:

- The value exists but is missing, meaning it is incomplete.
- The value does exist, which does not count as incomplete.
- The value may exist, but it is not actually known whether it exists or not.

Consistency is the third DQ dimension cited. Data stored must not have conflict between then [**28**], it must not violate semantic rules [**22**]. The result of a COVID-19 test in each the outcome is either positive, negative or inconclusive, should always be declared with these values. Similarly, the percentage of consistency can also be calculated by dividing the "number of consistency value" and "total number of the value", [**28**].

Another highlighted dimension is *Currency*, also known as *timeliness*, which refers to the fact that stored data must be sufficiently up-to-date [**23**]. Refers to how up-to-date the data is concerning the time it is accessed or used [**21**]. Data regarding the events should be updated in the systems as soon as possible. For instance, if a patient takes an ECG exam that results in some diagnosis and later take another ECG exam that leads to a different diagnosis, this should be updated in the systems as soon as possible for all the involved person to have the most recent diagnoses regarding that patient.

The *Usability* dimension is related to the understanding and accessibility of the available data. In the healthcare context, a lot of non-structured data (free text notes) used to be available, making interpreting this data more difficult. Additionally, because of a busy schedule, events in the healthcare settings or even the patients' data may not be completed timely, resulting in unusable data in the feature. In addition, abstract documentation regarding a patient diagnosis can result in poor data utilisation. In opposition, many details can add extra complexity to the data, resulting in a lack of usability [**6**].

Another indicated dimension is the *Relevance*; data must represent current or potential analysis needs. This dimension may be a challenge to guarantee due to the lack of awareness of the type of analysis needed when constructing a healthcare information system. Lastly, the *Duplication* dimension is also pointed. According to [**6**], duplication is defined as the presence of multiple records regarding the same entity and is often caused by the lack of relationship between data sets and poor systems integration. For instance, the presence of multiple registers for the same patient can cause some confusion, unsure which record contains the most recent and accurate information about the patient's condition and treatment. .

2.2.2. Methodology and framework for data quality evaluation and monitoring

Having in mind the process of data quality evaluation, different methodologies and frameworks have been proposed by various authors to perform such evaluation.

In [**21**], a data quality assessment methodology called *Complete Data Quality Methodology (CDQM)* is presented. It focuses on assessing and improving data quality and can be applied to structured relational data and semi-structured information. The methodology is composed of 8 steps.

The first step, known as *Reconstruct the State of Data*, refers to the process of analysing and understanding the current condition and characteristics of data. It involves assessing various aspects of data quality, such as accuracy, completeness, consistency, and correctness. This process often involves using methodologies and techniques, such as data cleaning, duplicate detection, object identification, and data quality assessment. The goal is to obtain a comprehensive and accurate understanding of the data, enabling effective data-driven decision-making and ensuring the reliability and usefulness of the data for various applications. After the *Reconstruct Business Processes* is performed, it refers to the process of identifying and understanding the various processes within an organisation and their relationships with different organisational units. This step involves determining the owner of each process and the units that participate in its execution [**21**].

The third step is *Reconstructing macro processes and rules* and involves analysing existing processes and rules to identify areas of inefficiency or ineffectiveness and then developing new processes and rules that better align with the organisation's or system's goals and objectives. Follows the step *Check Problems with Users*, which involves identifying any issues or problems that users may be experiencing in terms of causes of poor data quality. In step 5, *Measure Data Quality*, relevant dimensions and measures will be selected to provide a quantitative evaluation of the system's state. The following steps, namely *Set New Target Data Quality Levels, Choose Improvement Activities, Choose Techniques for Data Activities, Find Improvement Processes and Choose the Optimal Improvement Process* aim to outline actions to promote the system's improvement by evolving the data activities and business processes of the organisation [**21**].

The *Data Quality Management Framework* is another framework suggested by [**27**] to assess and monitor the data quality in the context of big data. This framework embodied the rule-based quality methodologies, highlighting its importance in data quality assessment. The authors affirm that since the concept of data quality is not universal and may vary depending on the application domain, to ensure data quality, a set of quality rules needs to be established, which includes constraints on data generation, entry, and creation. These rules can be created or discovered to correct or eliminate this data's poor data quality. However, the rules themselves are only one part of the data quality assessment approach. The proposed framework comprises ten key components, namely:

(1) *Exploratory Quality Profiling* involves analysing the data to understand its characteristics and identify potential quality issues. It is also responsible for automatically exploring data quality dimensions. It generates a list of quality rules proposals based on the evaluation of key data quality dimensions such as completeness, accuracy, and uniqueness;
(2) *Quality Mapping and Selection* responsible for mapping data features or attributes to Data Quality Dimensions and selecting appropriate metrics for evaluation;
(3) *Quantitative Quality Evaluation* involves assessing data quality based on user requirements and utilising a set of metrics to evaluate the attributes for a given set of Data Quality Dimensions;
(4) *Quality control* that involves the continuous monitoring and validation of the Data Quality Profile;
(5) *Quality Rules Discovery* that involves the automatic identification and formulation of data quality rules, which helps in detecting inconsistencies and anomalies in the data and consequently contribute to ensuring data quality;
(6) *Quality rules validation* ensure the accuracy and effectiveness of the identified quality rules. The validation process involves checking the validity and applicability of the quality rules in the Data Quality Profile configuration. The validation is performed based on the quality requirements and criteria defined for each Data Quality Dimension;
(7) *Quality Rules Optimization* aims to improve the efficiency and effectiveness of the quality rules applied to the data;
(8) *Data Quality Monitoring* that involves continuous quality control processes where quality reports are generated during each quality monitoring iteration on the data sets from the data source and later added to the data quality profile, updating it frequently;
(9) *Data processing, analytic, and visualisation* involves the application of algorithms or methodologies to extract insights from the available data, ensuring enhanced data quality, analysing the processed data to uncover patterns, trends, and valuable insights that can drive data-driven decision-making and projects the value

of processed data visually through dashboards and graphically enhanced charts, making it easier for decision-makers to understand and act upon the insights.

By implementing this framework, organisations can improve data quality, detect and correct any data quality management failures, and ultimately support data-driven decision-making processes [**27**].

Another study conducted by [**29**] proposed a framework to evaluate the quality of big data that considers DQ dimensions and weighted metrics. They affirm that since the importance of the information contained within data may vary according to the business point of view, more relevant data must have a higher impact when measuring the data quality. A framework was proposed where the quality of a data set is calculated by attributing a weight to the fields in analysis to calculate the percentage of success of a specific dimension or to the dimension itself to calculate the the data quality score of the data set. The weight represents the relevance of the field or dimension in data quality measurement. They took as an example a data set of customers in which it was first intended to calculate the percentage of completeness. The data set was composed of nine fields, namely, email, phone number, address, city, country, first name, last name and age. The percentage of completeness of each attribute was calculated by taking the average of the completeness of all attributes. The outcome was 60.62%. However, when calculated considering the weight of each field, where each weight was attributed according to the relevance of the field for the business purpose, the outcome was only 45.5%. The same approach was used to calculate the final scored data quality for the data set, where different data quality dimensions were measured, and each was attributed a weight according to the relevance they had to the business context. Dimensions like security, integrity and completeness were attributed a higher weight, meaning that they have greater relevance in these business contexts and consequently will have a greater impact on the final score data quality calculation.

By exploring different literature in the data quality field, it could be noticed that ensuring data quality can be challenging for many reasons. When applied in the healthcare context, the difficulty can increase, taking into account the complexity of the subject itself. Nevertheless, different frameworks and methodologies were proposed to assist in such a complex task. When mentioning frameworks for data quality assessment, all present one common aspect, the data quality dimension, meaning that it represents the core factor when considering evaluating the quality of a given data set.

The explored literature allows us to take essential insights that will support the development of this academic work. The proposed methodology for the practical use case of this thesis incorporates the definition and validation of the quality rules cited by [**27**]. These rules allow for identifying possible problems related to poor data quality. Allied with the rules, the methodology includes the concept of weighted metrics presented by [**29**]. Since it is understood that not all data have the same importance for an organisation and subsequently it shouldn't have the same impact when evaluating the data quality, the

attribution of weight for each quality rule allows to differentiate the level of importance of each one. Lastly, considering the importance of DQ dimension in the DQ evaluation process, which different authors have pointed out, the proposed methodology also incorporates the evaluation of the data, taking into account the data quality dimensions as one of the main steps of the evaluation process.

CHAPTER 3

Santa Maria Hospital case study

As previously mentioned, the main goal of this work is to evaluate the quality of a subset of data provided by a Portuguese hospital so it could be integrated into a common shared repository within the scope of a European project. So, this chapter addresses the steps needed to perform such an analysis.

Firstly, data profiling is conducted in the source database to understand the main characteristics of these data. Then, the main issues of these data are identified and mapped with the respective DQ dimensions. Lastly, a set of quality rules is defined, serving as a guide to address the listed issues and avoid the same ones in future collected data.

3.1. Context

Capacity is a project that aims to collect data regarding the cardiovascular history, diagnostic information and occurrence of cardiovascular complications in COVID-19 patients from different centres in a standardised manner. This will aid in providing meaningful insights regarding the incidence of cardiovascular complications in patients with COVID-19 and the vulnerability and clinical course of COVID-19 with underlying cardiovascular disease. Since the beginning of the project, different centres from all over the world have been given their contributions, including centres of Portugal, namely Hospital Espírito Santo Hospital and Hospital Prof. Doutor Fernando Fonseca [**9**].

Hospital Santa Maria is a public hospital centre in Portugal, inaugurated in April 1953 and is part of the university hospital centre in northern Lisbon. With a capacity of 1475 beds until February 2021, the hospital is the biggest in the country and has played a crucial role during the COVID-19 pandemic. Due to its capacity and the amount of data collected daily, the centre aims to collaborate with the Capacity project by providing the necessary data.

In this context, an integration between the Santa Maria hospital database and the Capacity database was proposed. Data from different sources intended to be analysed and migrated to the Capacity database, considering all the necessary measures to ensure that all the target database requirements would be fulfilled. During the migration process, one of the main steps to ensure the operation's success is a deep analysis of the data made available. The source database's main issues were identified to advise the best strategies to treat those issues before the data is loaded into the target database. To conduct this

analysis, the *ydata-profiling tool* [1] was used, and some issues were identified regarding several source database tables.

To develop this work, the Santa Maria Hospital provided 138 CSV files with data from different domains, such as patient admission, patient personal data, diagnosis, treatments, laboratory tests and results and so on. These data were imported to an SQL server database, resulting in 138 tables. The database was kept in a private server, with limited access to ensure the data's security. All sensitive patient data was anonymised, guaranteeing the patient's privacy. To access the data, it was mandatory to assign a confidentiality agreement to ensure that these data would only be used for this study purpose.

3.2. Data profiling analysis

Data Profiling is one of the first steps that should be done upfront in data integration projects, especially before applying transformation/cleaning procedures. Data Profiling techniques help to diagnose whether the data can meet the level standards for the purposes defined for the repository in which it will be stored [**30**].

To perform this analysis, the *ydata-profiling* was used, an open-source data analysis and manipulation tool built on top of the Python programming language [**31**]. Figure 3.1 illustrate the necessary steps to configure and use the Panda Profile.

First of all, it was necessary to install it in the environment where the analysis took place, using the command "pip install data-profiling". After the installation, the database connection was established, thus gaining access to all the tables available in the connected database. From this moment, the needed table was accessed, and the respective data profile report was generated in HTML format. The generated report includes a wide range of statistics and visualisations and a detailed analysis regarding each variable of the table under analysis, including information such as missing data, duplicate entries and outliers, variable correlations and so on.

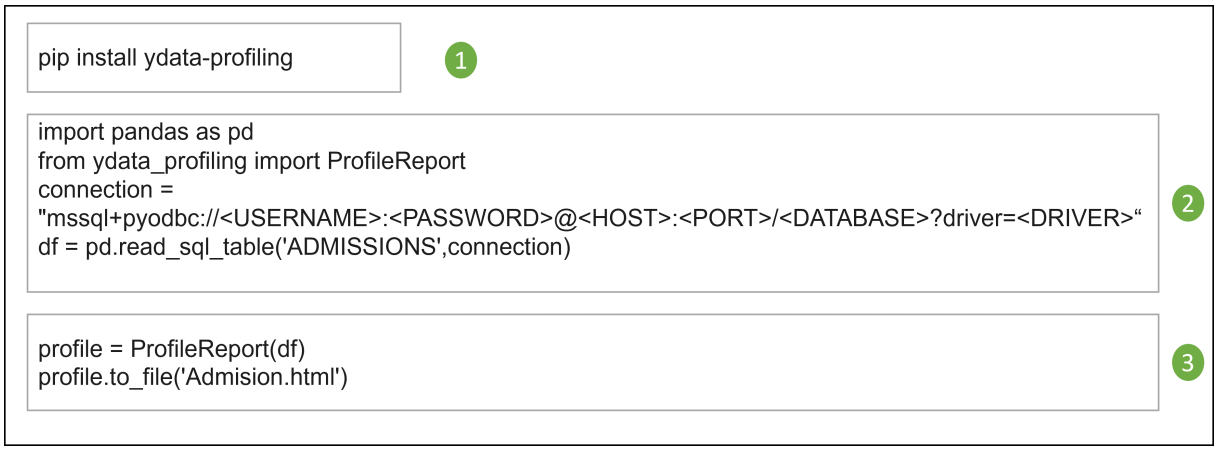


FIGURE 3.1. Configuration and usage of Panda Profile Tool

To get started with the data profile analysis, it was essential to get to know both the Capacity and Santa Maria Hospital databases.

[1] https://docs.profiling.ydata.ai/4.6/

The Capacity database aims to maintain records of patients with cardiovascular history, diagnostic information and occurrence of cardiovascular complications in COVID-19 patients. It comprises 14 tables with several attributes, where nine must be populated. For this academic work, only the nine mandatory tables are considered.

To get to know the source database, it was necessary to analyse and identify which tables would be needed to populate the Capacity database. From this analysis, 21 tables were selected from the source database, 17 of which are the target of a deeper analysis. of the remaining 4 tables were only used primary keys for join purposes.With the necessary tables selected, it was also crucial to identify which attributes would be used to populate the Capacity database, as it is more relevant that the analysis also focuses on these attributes. A brief description of each of the nine tables will be presented below, identifying the main tables and attributes necessary to populate each. The source table where no attribute was identified is because only the primary key was used to join the purpose:

(1) *Inclusion Criteria*, contains data that determine if this candidate meets the requirements to be considered as a participant. Table 3.1 presents the main tables used to populate it and the needed attributes.

(2) *Demographics*, stores data regarding the demographic information of the patients. Table 3.2 identifies the main tables used to populate it.

(3) *Cardiac Baseline Assessment*, have data regarding the hospital admission of the patient's vital signs, laboratory tests performed at the admission of the patient in the hospital, patient risk factors, current medication and cardiac problems. Table 3.3 presents the main tables used to populate it and the needed attributes.

(4) *Cardiac Biomarkers (Required – If Measured)*, stores data regarding all the cardiac biomarkers, namely the "Cardiac Troponin", "Creatine kinase (CK)", "CK-MB", "(NT-proBNP)BNP" when measured. On the table 3.4, we can find the used table to populate it.

(5) *Cardiac and Thromboembólic COVID-19 Complications*, stores data regarding cardiac or thromboembolic during hospitalisation after diagnosis with COVID-19. Table 3.5 presents the main tables used to populate it and the needed attributes.

(6) *Cardiac Outcome: 7 days follow-up*, contains data about the state of the patients seven days after the admission to verify if the cardiology symptoms are still involved. The main tables and attributes are in table 3.6.

(7) *Cardiac Outcome: 30 days follow-up*, stores data about the state of the patients 30 days after the admission to verify if the cardiology symptoms are still involved. The main tables and attributes involved can be consulted in Figure 3.6.

(8) *Discharge* contains data regarding the discharge of the patients either for death, palliative care, transfer to another facility, or recovery. Table 3.7 presents the main tables used to populate it and the needed attributes.

Table 3.1: Source tables used to populate ”Inclusion Criteria” table

Inclusion Criteria Table	
Source Table	Attribute
LABRESULTS	dboid textvalue observationtime
DEPARTMENTS	deptdboid deptdesc
COUNTRIES	countryboid countrydesc
PARTS	partboid partdesc
PICSDATA; ENVIRONMENTS; EMVIRONMENTSLOCATION; LOCATION; LABTEST; PARTCOMPONENT; COMPONENTS;ADMISSION; PATIENT	

Table 3.2: Source tables used to populate ”Demographic” table

Demographic Table	
Source Table	Attribute
SEXES	dboid textvalue observationtime
PATIENTS	patientdboid birthdate ethnicitydboid ehtnicgroupboid
ETHNICITIES	ethnicitydboid ethnicitydesc

ETHNICGROUPS	ehtnicgroupboid ehtnicgroupdesc
PICSDATA; PATIENT	

Table 3.3: Source tables used to populate "Cardiac Baseline Assessment" table

Cardiac Baseline Assessment Table	
Source Table	Attribute
COMPONENTS	componentboid componentdesc
ADMISSIONS	admissiondboid hospitalstarted height weight
RTTDAT	Rttdatadboid; Strated; Picsdatadboid; CD06; CD57; CD50; C980; C013; C010; C981; C014; C100; C080; C120; C001
LABRESULTS	dboid textvalue observationtime
UNITS	unitdboid unitdesc
PICSDATA; ENVIRONMENTS; EMVIRONMENTSLOCATION; LOCANTION;DEPARTMENT; PARTCOMPONENT;PATIENT, LABTESTS	

Table 3.4: Source tables used to populate "Cardiac Biomarkers" table

Cardiac Biomarkers Table	
Source Table	Attribute

DEPARTMENTS	deptdboid deptdesc
UNITS	unitboid unitsymbol
COMPONNETS	componentdboid componentdesc
LABRESULTS	dboid textvalue observationtime
PICSDATA; ENVIRONMENTS; EMVIRONMENTSLOCATION; LOCANTION; PARTCOMPONENT;PATIENT, LABTESTS, ADMISSIONS, ENCOUNTERS	

Table 3.5: Source tables used to populate ”Cardiac and Thromboembolic COVID-19 Complications” table

Cardiac and Thromboembolic COVID-19 Complications Table	
Source Table	Attribute
H_DIAGNOSYS	diagdboid diagdesc
ADMISSIONDIAGNOSYS	admissiondboid diagdboid admdiagdate
ADMISSION; PATIENT	

Table 3.6: Source tables used to populate ”Cardiac Outcome (7 and 30 days) ” table

Cardiac Outcome (7 and 30 days) Tables	
Source Table	Attribute

H_DIAGNOSYS	diagdboid diagdesc
ADMISSIONDIAGNOSYS	admissiondboid diagdboid admdiagdate
ADMISSION; PATIENT	

Table 3.7: Source tables used to populate "Discharge" table

Discharge Table	
Source Table	Attribute
DISCHARGES	deschdboid deschdesc
ADMISSIONS	admissiondboid hospitalended
DIAGNOSES	diagdboid diagdesc
TREATMENTS	treatmentdboid brandname genericname
ORDERS	orderdboid initdate lastdate
PATIENTS; PICSDATA	

3.2.1. Source tables overview and main characteristics

Having identified the tables and attributes needed, an analysis took place, giving us an overview of these tables, describing them, as well as their corresponding attributes and highlighting the main characteristics of each one, such as the data type of each attribute, the percentage of null and the percentage of duplication. Note that for the attribute in which the percentage of duplication is described as NA, the duplicate value is expected to be a normal scenario. The description of each table can be found below.

ADMISSION - The admission table is responsible for keeping all the records regarding a patient's admission to the hospital. Information like admission date, weight, height, patient number, discharge data, and more are registered.

From a broader analysis, we noticed that in a total of 38 attributes, 16 of them have a very high level of missing values, more than 77%, which allow us to affirm that in most of the admission made in this hospital, 42% of information is not collected. Also, we noticed that 12 in 38 attributes is 99% of the case filled with a constant value, which brings no helpful information regarding that admission record since the value registered does not represent actual fact. However, from the 38 columns existing, only 10 of them are relevant for the Capacity database population and detailed information is presented in table 3.8.

Analysing the table 3.8, we can notice that 50% of the attributes have more than 73% of null values, which reaffirms the fact that a lot of information requested for Capacity population coming from this table is not available. Regarding a particular attribute, *countrydboid*, the number of duplicate values called attention. The attribute has 99.9% of duplicate value. Analysing that value allows us to identify that, in most cases, this attribute is filled with a value corresponding to an unknown country, meaning no useful information.

Table 3.8: Data profiling analysis of the ADMISSION table

Attribute Name	Description	Data Type	% of null	% of duplicate value
admissiondboid	PK of the ADMISSION	numeric (21, 0)	0	0
hospitalstarted	Day on which the patient was admitted to hospital	datetime	16.1%	NA
hospitalended	Day on which the patient left hospital	datetime	73.0%	NA
started	Day on which the patient was admitted to the PICS system	datetime	76.7%	NA
ended	Day on which the patient left the PICS system	datetime	76.7%	NA
weight	patient weight	numeric (12, 5)	84.4%	NA
height	Patient height	numeric (12, 5)	93.9%	NA
patientdboid	FK of the Patient table	numeric (21, 0)	0%	0%

countrydboid	FK to the Country table	numeric (21, 0)	0%	99.9%
dischdboid	FK of the Discharge table	numeric (21, 0)	0%	0%

COUNTRIES - The table countries hold records regarding the countries in the world, and it is usually associated with the admission table to identify from what country the admitted patients are. It has a total of 68 records and three attributes, of which two of them are relevant for the Capacity database, namely the unique identification of the table, *countrydboid* and the description of the table, *countrydesc*, as shown in table 3.9. For both attributes, was not identified null or duplicate values.

Table 3.9: Data profiling analysis of the COUNTRIES table

Attribute Name	Description	Data Type	% of null	% of duplicate value
countrydboid	PK of the country	Numeric(21,0)	0%	0%
countrydesc	Country name	varchar (128)	0%	0%

COMPONENTS - This table contains the individual items tested as part of a laboratory test (e.g., Troponin). It holds a total of 3156 records and is composed of 7 attributes, of which 55.7% are not filled. However, for the Capacity database, only two attributes are relevant, and they are described in the table 3.10. Through this table, we can also notice that, for the needed attributes, there is no null or duplicate value.

Table 3.10: Data profiling analysis of the COMPONENTS table

Attribute Name	Description	Data Type	% of null	% of duplicate value
componentdboid	PK of the Component	Numeric (21,0)	0%	0%
componentdesc	Component description	varchar (128)	0%	0%

DISCHARGE - The discharge table contains all possible causes for each patient being discharged from the hospital. This can be because it was transferred to another facility or another unit, was recovered or died. The table has just three variables, one representing the primary key of the table, the other the description and the last one the delete flag. The table has a total of 620 records, but 50% of the records are duplicate values, including the primary key. The attributes considered for the target database are presented in 3.11.

Table 3.11: Data profiling analysis of the DISCHARGE table

Attribute Name	Description	Data Type	% of null	% of duplicate value
dischdboid	PK of the Discharge	Numeric (21,0)	0%	50%
dischdesc	Description of the Discharge	Varchar	0%	52.1%

DEPARTMENT - The department table holds all the existing departments of the hospital. It has a total of 62 records, and 50% of those records are duplicate ones, including the primary key. With a total of 6 attributes, allow us to have information regarding the department, such as its name and the facility to which it belongs. Information like the department type is 90% of empty. In the Capacity context, only two attributes are made important, namely the department identification and the department description and can be found described in table 3.12.

Table 3.12: Data profiling analysis of the DEPARTMENT table

Attribute Name	Description	Data Type	% of null	% of duplicate value
deptdboid	PK of the Department	Numeric (21,0)	0%	58.1%
deptdesc	Description of the Department	Varchar	0%	53.2%

ETHNICGROUPS - This table contain records regarding all the possible ethnic group to which a patient can belong. With three columns and nine records, the tables have no null value or duplicate value for both the attributes used in the Capacity database context. Table 3.13 represent that information, as well as the description and data type of each attribute.

Table 3.13: Data profiling analysis of the ETHNICGROUPS table

Attribute Name	Description	Data Type	% of null	% of duplicate value
ethnicgroupdboid	PK of the ETHNICGROUPS	Numeric (21,0)	0%	0%
ethnicgroupdesc	Description of the ETHNICGROUPS	Varchar	0%	0%

ETHNICITIES - Similar to the ETHNICGROUPS table, this also carries data regarding the possible ethnicity of a given patient. Has a total of 10 records and three attributes. The ones used in the Capacity database context are described in table 3.14. None of the variables are null or duplicate values.

Table 3.14: Data profiling analysis of the ETHNICITIES table

Attribute Name	Description	Data Type	% of null	% of duplicate value
ethnicitydboid	PK of the ETHNICITY	Numeric (21,0)	0%	0%
ethnicitydesc	Description of the ETHNICITY	Varchar	0%	0%

H_DIAGNOSIS - The diagnosis table holds all the records regarding the possible diagnosis of the patients. It has a total of 108934 records and six attributes, where 3 of them are numeric, two are categoric, and one is a boolean type. Information like the diagnosis codes and their respective description can be obtained through this table. Table 3.15 describes the attributes used for the Capacity database population, as well the percentage of null and duplicate values for each one.

Table 3.15: Data profiling analysis of the H_DIAGNOSIS table

Attribute Name	Description	Data Type	% of null	% of duplicate value
disgdboid	PK of the DIAGNOSIS	Numeric (21,0)	0%	0%
diagdesc	Description of possible diagnosis for patients	Varchar	0%	$< 0.1\%$
diagcode	Code of the diagnosys	Varchar	$< 0.1\%$	0.1%

LABRESULT - This table holds the results of all the laboratory tests taken for the patients. It has a total of 11472316 records, values filtered from March 2020, when the first COVID-19 cases appeared in Portugal, and seventeen attributes. The exam results are saved in two specific attributes, one for the result in integer and the other for the text version of the integer. Through this table, it is also possible to identify of each exam the result belongs and the lower and height values possible for each exam. For the Capacity database population, only three attributes were relevant, and it is presented in 3.16. All the used attribute has a very low percentage of null value, no more than 0.1%.

Table 3.16: Data profiling analysis of the LABRESULT S table

Attribute Name	Description	Data Type	% of null	% of duplicate value
dboid	PK of the LabResult	Numeric (21,0)	0%	0%
textvalue	Outcome of the lab exam	Varchar	0.1%	NA
observationtime	Time of the lab exam outcome	DataTime	0.1%	NA

ORDERS - the orders contains all record regarding the treatment ordered to the patients. With a total of 1173346, the table is composed of 39 attributes. From these attributes, only 3 of them are used in the Capacity population, *orderdboid*, representing the unique identification of the record, *initdate*, representing the data when the treatment should be initiated and *lastdate* indicating the date when the treatment should be ended. The table 3.17 listed these attributes.

Table 3.17: Data profiling analysis of the ORDERS table

Attribute Name	Description	Data Type	% of null	% of duplicate value
orderdboid	PK of the Orders table	Numeric (21,0)	0 %	0%
initdate	Date which the treatment ordered should be initiated	Varchar	0 %	NA
lastdate	Date which the treatment ordered should be ended	DataTime	0 %	NA

PATIENT - The Patient table holds all the personal information regarding the patient of that facility. It has a total of 512764 records and is composed of 27 variables. Similarly, with the Admission table, a high percentage of the variables have a very high number of missing values, 38%, and some variables are filled with the exact same random value, for instance, variables like name, last name, middle name, what does not bring any useful information. Meanwhile, for the population of the target database, only a few attributes proved interesting, and they will be described in table 3.18.

Table 3.18: Data profiling analysis of the PATIENT table

Attribute Name	Description	Data Type	% of null	% of duplicate value
patientdboid	PK of the Patient	Numeric (21,0)	0%	0%
birthdate	BirthDate of the patient	Date	2%	76.2%
ethnicgroupdboid	FK for ETHNICGROUPS table	Numeric (21,0)	57.4%	NA
SEXDBOID	FK for Sex table	Numeric (21,0)	0%	NA
ethnicitydboid	FK for ETHNICITIES table	Numeric (21,0)	57.6%	NA

PARTS - Parts table contains information on the type of laboratory analysis performed (e.g., "SARS-CoV-2 IgG IgM serology"). Holding 1735 records and composed of 4 attributes, only two were important in the Capacity population context. The table 3.19 presents the description of each variable as well as the percentage of null and duplicate values. As noted, both the variables have no record of null or duplicate values.

Table 3.19: Data profiling analysis of the PARTS table

Attribute Name	Description	Data Type	% of null	% of duplicate value
partdboid	PK of Part	Numeric (21,0)	0%	0%
patrdesc	Part description of the patient	varchar (128)	0%	0%

RTDATA - This table contains the data related to the physiological variables of the patients (e.g., Temperature, Blood Pressure). Holding a total of 30404477 records and 664 attributes, only 16 of them will be analysed, the ones used in the Capacity database population. Variable regarding the temperature of the patient, blood pressure, respiratory rate, oxygen saturation and heart rate of each patient is registered in this table when measured. Different attributes are used to register the same physiological variables, like in the case of C057 and C050, which are both used to hold the temperature of patients, and C100 and C011, which are also both used to keep the respiratory rate of the patients and so on.

The table 3.20 describe all the used variable. As we can notice in the referred table, information like temperature and blood pressure have a very high percentage of null value,

more than 97%, meaning that this information is not collected frequently or is not registered in the system. Analysing attributes regarding the respiratory rate, the percentage of null decreased to approximately 55%, meaning that in almost half of the admissions made, these data are registered in the system. Regarding the information about oxygen saturation and heart rate, they are registered in 90% and 94% of the admissions made, respectively, which represents a good rate in comparison to the other variables.

Table 3.20: Data profiling analysis of the RTTDATA table

Attribute Name	Description	Data Type	% of null	% of duplicate value
rttdatadboid	PK of the RTDATA	Numeric (21,0)	0%	0%
started	Datetime the physiologic variable was read	datetime	0%	0%
picsdata	FK to PICISDATA	numeric (21, 0)	0%	0%
cd06	Temperature (Manual Registration)	numeric (8, 3)	97.9%	NA
c057	Temperature	numeric (8, 3)	98.3%	NA
c050	Temperature	numeric (8, 3)	98%	NA
c980	Systolic Blood Pressure (Art.) 2	numeric (8, 3)	99.8%	NA
c013	Systolic Blood Pressure	numeric (8, 3)	98.6%	NA
c010	Systolic Blood Pressure – Invasive.	numeric (8, 3)	10.7%	NA
c981	Diastolic Pressure (Art.) 2	numeric (8, 3)	99.8	NA
c014	Diastolic Blood Pressure	numeric (8, 3)	98.6%	NA
c011	Diastolic Blood Pressure – Invasive	numeric (8, 3)	10.7%	NA
c100	Respiratory Rate	numeric (8, 3)	49.4%	NA
c080	Respiratory Rate	numeric (8, 3)	55%	NA
c120	Oxygen saturation (SpO2)	numeric (8, 3)	9.3%	NA

c001	Heart Rate	numeric (8, 3)	5.8%	NA

SEXES - The sexes table holds the record regarding the sex of the patient. It has a total of 8 records and three attributes, two of which are used in the Capacity database population. The main characteristics of these attributes are described in table 3.21. We can notice that there are neither null nor duplicate values from both attributes.

Table 3.21: Data profiling analysis of the SEXES table

Attribute Name	Description	Data Type	% of null	% of duplicate value
sexdboid	PK of the SEX	Numeric (21,0)	0%	0%
sexdesc	Description of the Sex	Varchar	0	0

TREATMENT - This table is responsible for keeping all the possible treatments that can be ordered for a given patient. With a total of 12223 records and 28 attributes, different information can be obtained through this table, such as the commercial name and generic name of the treatment, the group of the treatment belonging, the family, it is a narcotic or antibiotic for instance, as well as if is an external treatment or internal one. Fourteen is null in 100% of the record, meaning that 50% of the attribute is not filled in a single record. In the Capacity context, only three attributes were used, and although it presents no null value, they have some percentage of duplicate value, as shown in the table 3.22.

Table 3.22: Data profiling analysis of the TREATMENT table

Attribute Name	Description	Data Type	% of null	% of duplicate value
treatmentdboid	PK of Treatment	Numeric (21,0)	0%	0%
brandname	Brand name of the treatment	Varchar	0%	1.4%
genericname	Generic name of the treatment	Varchar	0%	16.6%

UNITS - This table keeps the information regarding the units of measurement used in the hospital (e.g., Grams). Composed of 9 attributes, it holds a total of 266 records. As listed in the table 3.23 for the Capacity database, only two attributes was necessary. The attribute ”UNITSYMBOL” has 15.3% of duplicate value, records whose importance can be reassessed since they represent the same unit.

Table 3.23: Data profiling analysis of the UNITS table

Attribute Name	Description	Data Type	% of null	% of duplicate value
unitdboid	PK of Units	Numeric (21,0)	0%	0%
unitisymbol	Abbreviation used for the unit of measure	varchar (16)	0%	15.3%

From the tables PICSDATA, ENVIRONMENTLOCATION, ENVIRONMENT, LABTESTS and LOCATIONS, only identifier attributes were utilised to cross-reference information between different tables. All the primary keys show up as being unique and with no null values and are numeric (21,0) data types.

3.2.2. Main issues and their respective DQ dimensions

At this point, each table and attribute's main characteristics were known. Afterwards, a deeper analysis was conducted to identify each table's main data quality issues. Ultimately, a map between these issues and the respective DQ dimension violated was performed. The issues identified have been categorised into four dimensions: *Accuracy*, *Completeness*, *Duplication* and *Consistence*.

According to [**21**], issues related to the *accuracy* dimension can be identified by checking whether v is any one of the values in D. Whenever the issues identified, following in the description that does not belong to a subset of data that was supposed to do, is categorised as an accuracy problem. *Duplication* occurs when a record regarding an entity is stored twice or more in a data source [**21**]. The issues identified in the source database where this happens are framed in this dimension. Regarding *consistency*, [**21**] also points out that this dimension is violated when semantic rules defined over data items are violated. Therefore, when the identified issues represent that kind of problem, it is categorised as a violation of this dimension. Lastly, an issue is considered a Completeness issue when it has null values when these values, in fact, exist or might be present, but their existence is uncertain [**21**]. Issues such as duplicate value, domain violation, high percentage of null value, primary key violation, misspelling error, heterogeneity of measure and so on were identified and listed in the table 3.24.

Table 3.24: Summary of the identified issues in the source database

Table	Attribute name	Issue
ADMISSIONS	countrydboid	(**ISS1**) 99.1% of the record has the same value that corresponds to an unknown country, which gives us no meaningful information
ADMISSIONS	weight	(**ISS2**) Huge percentage of null value - 84.4% of the record has no information regarding the weight of the patient (**ISS3**) Domain violation – existence of record which negative value (-1) (**ISS4**) Presence of outliers - patient registered with a weight of over 400 kg
ADMISSIONS	height	(**ISS5**) Huge percentage of null value - 93.9% of the record has no information regarding the height of the patient (**ISS6**) Heterogeneity of measure units – the height of the patient is sometimes registered in centimetres and other times in meters, (ex: 1.70 and 175.0) (**ISS7**) Domain violation – existence of record with negative height (-1)
ADMISSIONS	dischdboid	(**ISS8**) For all the patient discharges, in 67% of the cases the reason why they were discharged is registered as an ”unknown” reason.
COMPONENTS	componentdesc	(**ISS9**) Different record representing the same component (ex: Resultado:..... and Resultado: .....)

DISCHARGE	dischdboid	(**ISS10**) Primary key violation – the existence of two records with the same primary key (**ISS11**) Duplicate row – the existence of two rows with exactly the same value in every single attribute of the database
DEPARTMENT	deptdboid	(**ISS12**) Primary key violation – the existence of two records with same primary key (**ISS13**) Duplicate row – the existence of two records with exactly the same value in every attribute of the database
DEPARTMENT	deptdesc	(**ISS14**) Misspelling error – all the records are written in Portuguese while a single one is right in English, namely “Preadmit Department”
ETHNICITY	ethinicitydesc	(**ISS15**) Domain violation – existence of record that does not represent any ethnicity (ex: ”Patient declined”)
H_DIAGNOSIS	diagdesc	(**ISS16**) Duplicate value – different records with the same description of the diagnosis (ex: two rows with “Leucemia ou Linfoma” description)
LABRESULT	textvalue	(**ISS17**) For the covid exam, the result is declared with different descriptions (“POSITIVO para Coronavírus (ex: SARS-CoV-2”, “POSITIVO”, “POSITIVO (> 12500 U/mL)”, “Resultado positivo ontem. Não se justifica repetição”) (**ISS18**) Domain violation - results declared with values that do not allow to know the final outcome of the test (ex: ”-”)

ORDERS	initdate	(**ISS19**)	Incoherent date - start date greater than end date
ORDERS	lastdate	(**ISS20**)	Incoherent date - end date less than end date
PATIENT	sexdboid	(**ISS21**)	Variable with 7 distinct values when in reality only 3 active records regarding sex existed (F, M and N). The record that was registered with old records ("Desconhecido", "Desconheci", "Feminino", "Masculino"), was not updated with the new ones created
PATIENT	birthdate	(**ISS22**) (**ISS23**)	1.3% of missing value Value out of range, (ex: 1779-11-13 and 2033). It is not possible to have a record of 1779 because the hospital didn't even exist and a birthday date in the future is also not correct
PATIENT	ethnicgroupdboid	(**ISS24**) (**ISS25**)	Huge percent of missing value (57.4%) For the remaining 42.6% that does not have a null value, 65% are associated to a unknown ethnic group.
PATIENT	ethnicitygroupdboid	(**ISS26**) (**ISS27**)	Huge percent of missing value (57.6%) For the remaining 42.4% that does not have a null value, 98.7% are associated to a unknown ethnicity.
PARTS	partdesc	(**ISS28**)	Different record representing the same part (ex: Resultado ..... and Result ....)
RTDATA	code	(**ISS29**)	Ambiguous violation – different codes representing the same concept, (ex: C057 and C050 represent the blood temperature; C100 and C80 represent respiratory rate)

TREATMENT	brandname	(**ISS30)**	Misspeling error – (ex: “regulacao” instead of “regulação”; “composicao” instead of “composição”; etc)

Each one of the identified issues in Table 3.24 represents a violation of some data quality dimension previously mentioned at the beginning of the chapter, and this mapping can be observed in Table 3.25. As we can observe, the DQ dimension with the most related issue is accuracy, followed by consistency, meaning that special attention should be taken regarding these dimension.

Table 3.25: Mapping between data quality dimension and identified issues

DQ Dimension	Related Issues
Completeness	(**ISS1**);(**ISS2**);(**ISS5**);(**ISS22**);(**ISS24**);(**ISS26**)
Accuracy	(**ISS3**);(**ISS4**);(**ISS16**); (**ISS8**); (**ISS14**); (**ISS15**); (**ISS18**); (**ISS21**); (**ISS23**); (**ISS25**); (**ISS27**); (**ISS30**)
Duplication	(**ISS11**);(**ISS13**);(**ISS16**)
Consistency	(**ISS6**);(**ISS9**);(**ISS10**);(**ISS12**);(**ISS17**); (**ISS19**);(**ISS20**);(**ISS28**); (**ISS29**)

3.3. Quality rules characterization based on profile analysis

To ensure that a set of data would fit some purpose, it is important to define quality rules that can later be validated using these data. These rules were identified after ”x-raying” a significant sample of the data. I was derived from a first analysis and reflected the desirable requirements that this data should fulfil, although some of them are violated. Having these rules defined would allow set measures to correct the current data to improve the quality and prevent future data from having the same issues, thus ensuring a better quality of the data collected. Since future data may have new requirements, these rules should always be adapted to accompany the evolution of business needs.

Having that in mind, in this section, a set of quality rules will be defined, taking into account the main tables and attributes previous identified to populate the target database. All these quality rules are listed in the table 3.26.

Table 3.26: Rule

Rule type	Description	Attribute name
Uniqueness	Data must be unique	admissiondboid countrydboid countrydesc componentdboid componentdesc dischdboid deptdboid deptdesc ethnicgroupdboid ethnicgroupdesc ethnicithdboid ethnicitydesc diagdboid diagcode dboi patientdboid
Null or Empty	Data must not be null neither empty	admissiondboid patientdboid countrydboid countrydesc componentdboid componentdesc dischdboid deptdboid deptdesc ethnicgroupdboid ethnicgroupdesc ethnicithdboid ethnicitydesc diagdboid diagcode dboi observationtime

Data Validation	If hospitalstarted is null, then started should not be null and should be less or equal to the current data; otherwise can be null; hospitalended must be larger or equal to hopitalstarted ended must be larger or equal to the started observationtime must not be lower than the hopitalstarted birthdate must be less than the current date	started hospitarstarted hospitalended ended observationtime birthdate
Data constraint	weight should be between 2.5 and 120 kg height should be between 48 and 250 cm sexdboid must represent the values feminine, masculine or others	weight height sexdboid
Data format	The measure unit for weight should be kg The measure unit for heigh should be cm	weight heigh
Data integrity	Data must belong to a valid entity	patientdboid countrydboid dischdboid componentdboid
Grammatical	Data must be described in a common language (only Portuguese or only English)	dischdesc deptdesc brandname genericname

CHAPTER 4

Data Quality Assessment

When we mention data quality assessment, different methodologies and frameworks can be used. According to [**21**], the choice of methodology must take into account a number of factors. For instance:

- *Scope and Objectives* since different methodologies may be more suitable for different purposes, such as assessing overall data quality, identifying specific data quality issues, or evaluating the impact of data quality on business processes;
- *Data Characteristics*, take into account the characteristics of the data being assessed, such as volume, variety and velocity, once some methodologies may be better suited for structured data, while others may be more effective for unstructured or semi-structured data;
- *Available resources and expertise*, including time, budget, and expertise, for implementing the chosen methodology, since some methodologies may require specialised tools or skills, while others may be more resource-friendly and easier to implement and so on.

In this academic work, the methodology used is an adaptation of two proposed methodologies of the authors [**27**] and [**29**], combining the concept of quality rules, where pre-defined rules are used as criteria to evaluate and assess the quality and the weight metrics, a concept that proposes to associate weight to the data as a way of highlighting the importance of one piece of data over another, respectively.

This chapter addresses the methodology used to assess a significant subset of the data of the Santa Maria Hospital. After is described the configuration of the used framework for the data validation, based on the quality rule proposed, Great Experience (GE) [1]. To perform the data assessment, five analytical questions are proposed. For each of them, a set of quality rules is defined and subsequently validated using the GE. Each rule is also mapped with the corresponding DQ dimension, and a weight is attributed, considering the rule's importance in the context of the proposed question. Lastly, for each question, a score is calculated based on the weight of the quality rule and the percentage of success resulting from each validation. This score represents the quality score of the subset of the data assessed in the context of each question.

4.1. Methodology description

The proposed methodology intends to use quality scores to evaluate data quality. The choice of such methodology is because the quality score provides a quantitative measure of

[1] https://greatexpectations.io/

data quality, allowing for easy comparison and assessment of different data sets. It helps identify the strengths and weaknesses of data quality by considering multiple metrics and DQ dimensions that can be later used to prioritise data cleaning and improvement efforts, focusing on areas with lower scores, enabling organisations to set benchmarks and goals for data quality improvement and tracking progress over time. [**29**]

To achieve such an objective, the methodology relied essentially on 2 points:

- Establish a set of quality rules that would be validated and serve as guidelines for future correction and elimination of poor data quality. Each rule falls under a certain DQ dimension and, when not successfully validated, represents a violation of such dimension, which would contribute negatively to the evaluation of the dimension;
- Apply weight metrics at two levels: quality level to calculate the score at DQ dimension level and DQ dimension level to calculate the overall data quality score. By applying data weights at different levels, a quality score can provide a more accurate assessment of data quality, considering the relative importance of different metrics and data fields.

Overall, using a quality score provides a standardised and comprehensive approach to evaluating data quality, enabling organisations to make informed decisions based on reliable and trustworthy data.

The methodology used to perform the DQ assessment compresses the following steps:

Step1 - Propose an analytical question, based on the main goals of the Capacity project.

Step2 - Identify the used tables and attributes to answer the proposed question

Step3 - For each table and their respective attribute, define the necessary quality rules and map them with their corresponding DQ dimension

Step4 - For each rule, validate it through GE, indicating the percentage of success of the validation. This represents the percentage of data that fulfils the requirements of the defined rule

Step5 - Group the defined rule by DQ dimension and assign for each of them a level of significance in a range of 1 to 10, where one corresponds to a less important rule and ten a more important rule. The table 4.1 was used as a guide to attribute the factor

Step6 - Calculate the weight of each field. The weight is the ratio of the assigned level to the total level of that dimension, and the sum of all of them will represent a total of 1

Step7 - Calculate the score of each DQ dimension by summing the product of each weight per the percentage of success validation of the corresponding rule

Step8 - Assign for each DQ dimension a level of significance 1 to 10 and use the same method described in Step 6 to calculate the weight for the dimensions

Step9 - Calculate the final score of the question by summing the product of each dimension weight per the score of the corresponding dimension previously calculated.

Table 4.1: Level of significance degree

Level of significance	1-2	3-4	5	6-7	8-10
Impact degree	Very Low	Low	Moderate	Significant	Very High

4.2. Configuration and usage of Great Expectation framework

To start using the GE framework,three essential steps were followed, as shown in figure 4.1, and each of them will be explained in detail below.

Step1 Installation of the framework - The first step to use the GE framework was to install it where the development took place. To do such an installation, the ”pip” command was used.

Step2 Once the installation was successfully completed, the connection with the *aimhealth_db* database, which holds the data of Santa Maria Hospital, was established and a SQL data source was created. For that was necessary to create a connection string identifying the database name, as well the username and the password to access and the drive that will be used.

Step3 Having created the data source, through the database connection, all the tables of the connected database were made available and could be used as table assets. This asset was subsequently used to build batches of data that were used to be validated. Expectations were created for each rule wanted to validate and afterwards saved and used. The final result could be obtained either in a JSON format or HTML report format to be analysed and evaluated.

A crucial step of the data validation through the GE framework is the configuration of the *Expectations*. Expectations are statements made regarding your data. These statements are conveyed using declarative language through straightforward and human-readable Python methods [**32**]. For example, in order to assert that you want the column “patient_age” to be between 0 and 110, you can say:

```
expect_column_values_to_be_between(
column=”PATIENT_AGE”,
min_value=1,
max_value=110
)
```

GE then uses this statement to validate whether the column patient_age in a given table is, in fact, between 0 and 110 and returns a success or failure result. Currently, the framework provides several built-in *Expectations*, with the possibility to write custom ones. For this academic work, the necessary *Expextations* was created based on the defined rules. Five types of expectation were used and can be consulted in Table 4.2. Also, in some cases, additional configuration was added for these expectations to match the role's needs in validation. For instance, by specifying a condition in the variable *row_condition*,

a multi-column validation is performed where one column will only be validate if a filter is applied in the data being validated based on the condition passed:

```
  expect_column_values_to_not_be_null(
column="ETHNICGROUPDBOID",
row_condition="ETHNICITYDBOID_treated
Is None",
condition_parser="pandas",
)
```

Rule type	Expectation	Description
Uniqueness	expect_column_values_to be_unique	Used to validate roles related to the uniqueness of the attributes.
Empty or null values	expect_column_values_to_not be_null	Used to validate roles for verifying if a value is empty or null
Regex rule	expect_column_values_to match_regex	Use to validate specific roles based on regex conditions. For example, since the result of COVID-19 is not standardised, to obtain if the result was positive or negative, the attribute holding the result must have the word positive or negative. For *that the regex "(?i).*POSITIVO*|.*NEGATIVO*"* can be validate
Define set of value	expect_column_values_to be_in_set	Used to validate roles based in a range condition; For instance, the *patientid* in the ADMISSION table must belong to a valid patient. That can be verified by checking if the id belongs to the set of values defined in the PATIENTS table.
Define the range of value	expect_column_value_lengths_to _be_between	Used to validate roles based on a range of values. For instance, the patient's age must be between 0 and 110

Table 4.2: Mapping between the expectations and the type of role

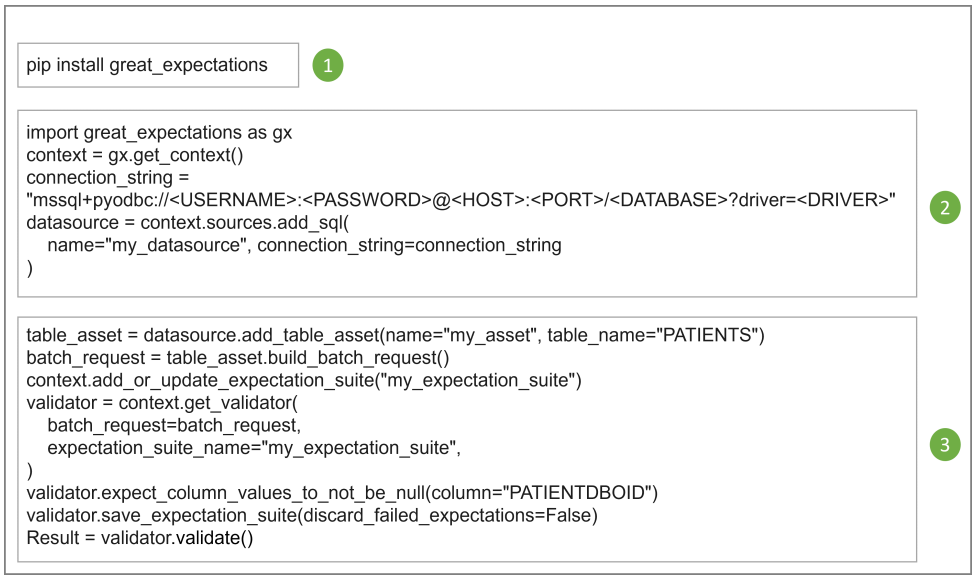


FIGURE 4.1. Configuration and usage of the GE framework

4.3. Implementation and validation of the quality rules

Previously, in section 3.3, generic quality rules were defined for the main tables used to populate the Capacity database. These rules can always vary, considering the business needs of the moment. A more specific evaluation was carried out to demonstrate such variation, and five analytical questions were suggested in the context of the Capacity database. The proposed questions are considered critical questions having in mind the target repository purposes. They will make it possible to diagnose the data quality without using irrelevant columns/tables, contributing positively or negatively to the final data quality assessment.

QN1 *Which patients are diagnosed with COVID-19?*

QN2 *What is the percentage of children (patients aged up to 12 years) diagnosed with COVID-19?*

QN3 Does the patient present any concern about cardiac biomarkers that could lead to heart failure, specifically Troponin or (NT-pro)BNP?

QN4 Which ethnic group detected the highest percentage of COVID-19 infection?

QN5 What percentage of patients diagnosed with COVID-19 die?

A DQ evaluation was performed for all the questions proposed to assess the confidence level in the answers based on the available data, considering the main attributes used to give such answers. A set of rules was defined and validated for each table and used attributes using GE. The corresponding DQ dimensions were also mapped, and weight was attributed to every quality rule, considering its relevance. The relevance of each rule is determined by considering whether it would be possible to answer the proposed question without ensuring it. The assigned weight will represent the level of importance of the rule/corresponding DQ dimensions in the context of that question, amounting to 100% when added together.

Since the information contained within the data does not have the same level of importance when used for different analytic proposals, the weight of each rule/corresponding DQ dimensions will also be determined according to the relevance of the rule and the DQ dimension to answer the question. Most relevant data must have a higher impact on data quality measurements and, therefore, greater weight.

An individual analysis of each question was conducted to find out the outcome of the assessment and is presented below.

(1) *Which patients are diagnosed with COVID-19?*

Having identified the analysis question, according to the first step of the methodology, the second step in performing the assessment was identifying the needed tables and attributes to answer these questions. As shown in Figure 4.2, to have the list of patients diagnosed with COVID-19, an inner join between 12 tables was made, namely ADMISSION, PATIENTS, PICSDATA, ENCOUNTERS, ENVIRONMENTS, ENVIRONMENTLOCATIONS, LOCATIONS, DEPARTMENT, PARTCOMPONENTS, and PARTS using the primary key and their respective foreign key. Besides the patient's identification number, attributes like *hospitalstarted* from the ADMISSION table, *departdesc* from DEPARTMENT table, and *textvalue* from LABRESULTS were also selected.

Afterwards, filters were performed to include only records that belonged to the department of Intensive Care Medicine Service (SMI), UCIMC or Respiratory Intensive Care Unit (UCIR), in which the data of admission in the hospital was greater than or equal to 02/03/2020, where the laboratory exams made corresponded to a COVID-19 exam and the final result was positive.

The final output was a total of 6385 patients diagnosed with COVID-19 in a universe of 512764 patients registered in Santa Maria Hospital, meaning that 1.3% of the patient registered in the hospital was diagnosed with COVID-19.

Afterwards, according to the third step of the methodology, needed rules were defined for each attribute and mapped with the corresponding DQ dimension, as shown in the 4.3. The quality rule defined fell into four categories, namely completeness, duplication, consistency and accuracy.

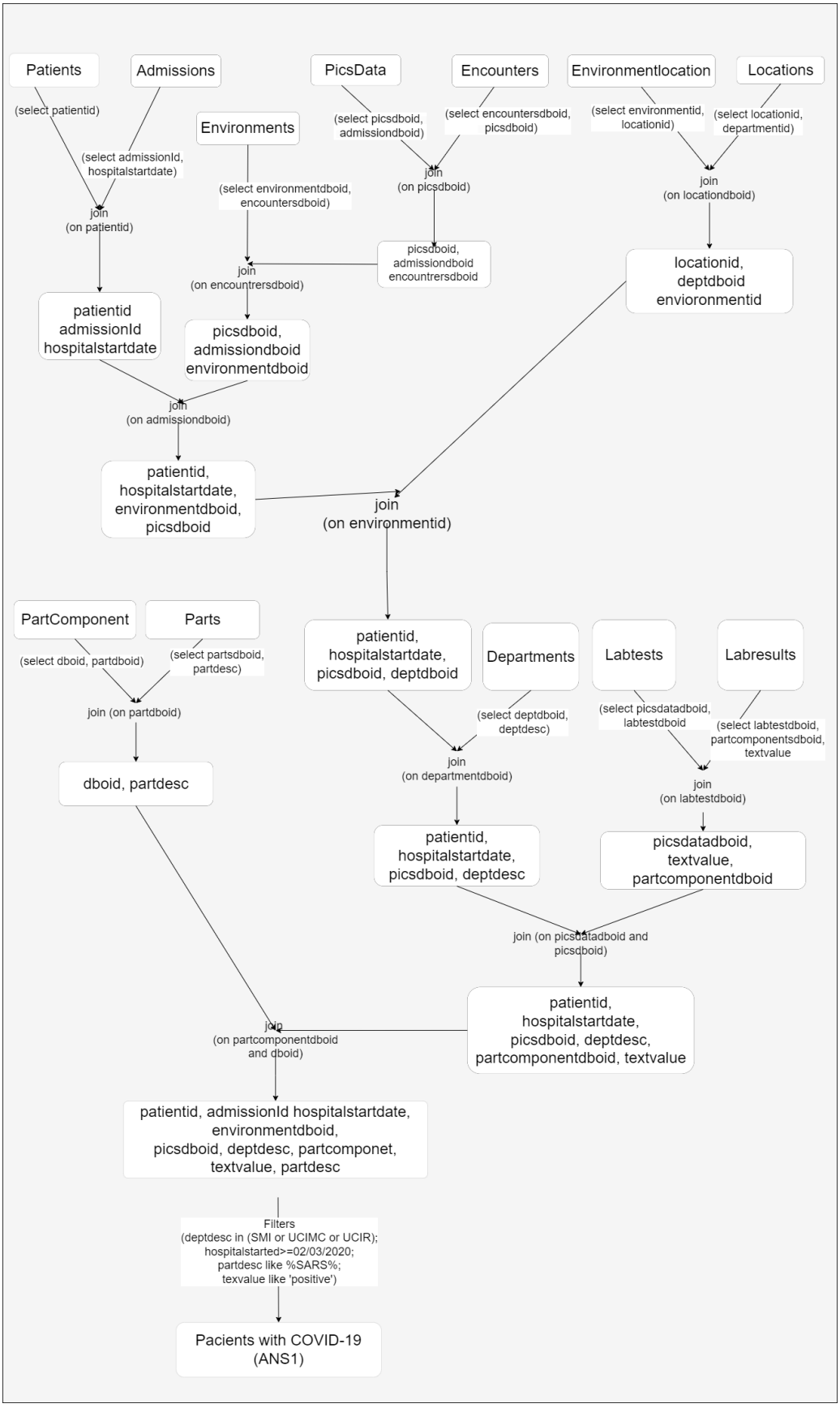


FIGURE 4.2. Tables and attributes used to identify patient with COVID-19 - QN1

Table 4.3: Quality rules defined to evaluate patients with COVID-19 - QN1

Table used	Attribute name	Quality rule		Dimension
ADMISSIONS	admissiondboid	**RL1**	admissiondboid must be unique	Duplication
		RL2	admissiondboid must not be null	Completeness
	patientdboid	**RL3**	patientdboid must not be null	Completeness
		RL4	patientdboid must belong to a valid patient	Accuracy
	hospitalstarted	**RL5**	hospitalstarted must not be null	Completeness
PATIENT	patientdboid	**RL6**	patientdboid must be unique	Duplication
		RL7	patientdboid must not be null	Completeness
LABRESULT	dboid	**RL8**	dboid must be unique	Duplication
		RL9	dboid must be not null	Completeness
	textvalue	**RL10**	for all result of COVID-19 test, textvalue must include ”POSITIVE” or ”NEGATIVE” word	Consistency

PARTS	partdboid	**RL11**	partdboid must be unique	Duplication
		RL12	partdboid must be not null	Completeness
	partdesc	**RL13**	partdesc must not be duplicated	Duplication
		RL14	partdesc must exist at least one record including ”SARS”	Accuracy
DEPARTMENT	deptboid	**RL15**	deptboid must be unique	Duplication
		RL16	deptboid must be not null	Completeness
	deptdesc	**RL17**	deptdesc must not be duplicated	Duplication
		RL18	deptdesc must be not null	Completeness

In step 4 of the methodology, each quality rule was converted to ”expectation”, listed in table 4.4 and subsequently validated. In the referred table, we can find the result of the validations, presenting the percentage of success and failure of each one. A more detailed analysis of the validation performed is presented below.

Table 4.4: List of defined expectations for rules of the question - QN1

Expectation	Corresponding RL	% expected	% unexpected
expect_column_values_to_be_unique	RL1	100%	0%
expect_column_values_to_not_be_null	RL2	100%	0%
expect_column_values_to_not_be_null	RL3	99.9%	0.1%
expect_column_values_to_be_in_set	RL4	100%	0%
expect_column_values_to_not_be_null	RL5	83.9%	16.1%
expect_column_values_to_be_unique	RL6	100%	0%
expect_column_values_to_not_be_null	RL7	100%	0%
expect_column_values_to_be_unique	RL8	100%	0%
expect_column_values_to_not_be_null	RL9	100%	0%
expect_column_values_to_match_regex	RL10	37.5%	62.5%
expect_column_values_to_be_unique	RL11	100%	0%
expect_column_values_to_not_be_null	RL12	100%	0%

expect_column_values_to_be_unique	RL13	100%	0%
expect_column_values_to_match_regex	RL14	100%	0%
expect_column_values_to_be_unique	RL15	100%	0%
expect_column_values_to_not_be_null	RL16	100%	0%
expect_column_values_to_be_unique	RL17	100%	0%
expect_column_values_to_not_be_null	RL18	100%	0%

The validation of the defined "expectations" resulted in the output presented in Figure 4.3. For question QN1, nine attributes were used, and 18 expectations were defined and validated, where 5 of these attributes were successfully satisfied, and 4 of them were not.

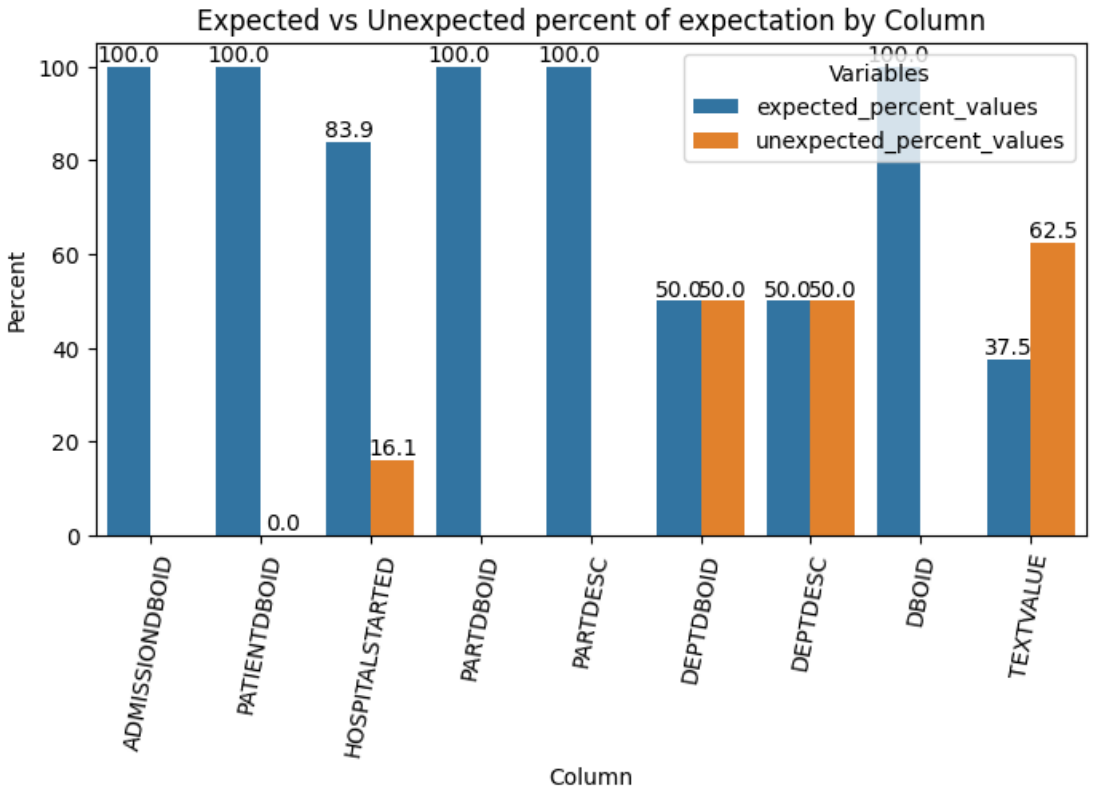


FIGURE 4.3. Expected vs Unexpected % of validated expectation by attribute - QN1

An in-depth analysis regarding the failed expectation shows that approximately 16% of the admissions made in the hospital did not register the entrance data. Regarding the department variables, both *deptdboid* and deptdesc only pass the validate expectations with a 50% score because these tables have 50% of duplicate records. Mentioning the *textvalue* variable, which holds the result where the patient has been detected with COVID-19 or not, we can observe that 62% of the data validated did not satisfy the defined rule, **RL10**, that's because the result present did not say if the test was positive or negative, containing others information instead. This fact can represent a negative scenario when identifying patients with COVID-19 since neither half of the patients subjected to the COVID-19

test had registered a conclusive result, showing a lack of standardisation when declaring the results of COVID-19 exams.

At this point, all the rules were validated, and the level of significance and the weight for each rule were calculated according to steps 4 and 5 of the methodology. As presented in table 4.5, each rule's significance level is defined on a scale of 1 to 10, referencing the table 4.1. The rules are grouped by DQ dimension, and the sum of the weight of each rule by dimension should be 1. The weight is obtained by calculating the ratio of the assigned level of significance to the total level of significance of that dimension, as demonstrated in Equation 4.1

$$\text{Weight of RL} = \text{level significance of the RL/total level significance of the rule for DQ dimension} \tag{4.1}$$

For instance, to calculate the weight of the rule **RL2**, which belongs to the completeness dimension, in which the assigned level of significance is 5, and the sum of the total level of significance for each rule of that dimension, level of significance of **RL2**, **RL3**, **RL5**, **RL7**, **RL9**, **RL12**, **RL16** and **RL18**, is 47, we have the following result:

$$\begin{aligned} \text{Weight of } \mathbf{RL2} &= 5 * (5 + 8 + 3 + 8 + 5 + 8 + 5 + 5) \\ &= 5/47 \\ &= 0.11 \end{aligned} \tag{4.2}$$

For the obtained result, we can notice that the rule **RL2**, as a weight of 0.11, meaning that it has an impact of 11% when calculating the DQ score of the dimension to which it belongs.

Table 4.5: Level of significance and weight of each rule of the question - QN1 - grouped by DQ dimension

Quality Rule	Dimension	Level of significance(LS)	Weight
RL4	Accuracy	10	0,5
RL14	Accuracy	10	0,5
RL2	Completeness	5	0,11
RL3	Completeness	8	0,17
RL5	Completeness	3	0,06
RL7	Completeness	8	0,17
RL9	Completeness	5	0,11
RL12	Completeness	8	0,17
RL16	Completeness	5	0,11
RL18	Completeness	5	0,11
RL10	Consistency	10	1
RL1	Duplication	5	0,12
RL6	Duplication	7	0,16

RL8	Duplication	5	0,12
RL11	Duplication	8	0,19
RL13	Duplication	8	0,19
RL15	Duplication	5	0,12
RL17	Duplication	5	0,12

Since we have the weight for each rule, we calculate each dimension's score by multiplying each rule's success percentage by their respective weight, Table 4.5 and Table 4.4, respectively, according to step 7. Consider the description of each variable for the equation:

R - Rule

DQD Score - Data Quality Dimension score;

RW - Rule weight;

SSP - Success Percentage

$$DQD \text{ Score} = RWR_1 \times SSPR_1 + RWR_2 \times SSPR_2 + \ldots + RWR_n \times SSPR_n \quad (4.3)$$

While the weight of the rule represents the importance of that rule in the context of the proposed question, the score of the dimensions represents the success rate of each dimension. For instance, for the *Completeness* dimension, the obtained score was 99%, meaning that based on the rules evaluated, the completeness rate is 99%. The other 1% of the data is missing. The final result for each dimension can be consulted below.

$$\begin{aligned} \text{Score Consistency} &= RWR_{18} \times SSPR_{18} \\ &= 1 \times 37.5 \\ &= 37.5\% \end{aligned} \quad (4.4)$$

$$\begin{aligned} \text{Score Accuracy} &= RWR_4 \times SSPR_4 + RWR_{14} \times SSPR_{14} \\ &= 0.5 \times 100 + 0.5 \times 100 \\ &= 100\% \end{aligned} \quad (4.5)$$

$$\begin{aligned} \text{Score Completeness} &= RWR_2 \times SSPR_2 + RWR_3 \times SSPR_3 + RWR_5 \times SSPR_5 \\ &\quad + RWR_7 \times SSPR_7 + RWR_9 \times SSPR_9 + RWR_{12} \times SSPR_{12} \\ &\quad + RWR_{16} \times SSPR_{16} + RWR_{18} \times SSPR_{18} \\ &= 0.1 \times 100 + 0.17 \times 99.9 + 0.06 \times 83.9 + 0.17 \times 100 \\ &\quad + 0.11 \times 100 + 0.17 \times 100 + 0.1 \times 100 + 0.11 \times 100 \\ &= 99\% \end{aligned} \quad (4.6)$$

$$\begin{aligned}\text{Score Duplication} &= RWR_1 \times SSPR_1 + RWR_6 \times SSPR_6 + RWR_8 \times SSPR_8 \\ &\quad + RWR_{11} \times SSPR_{11} + RWR_{13} \times SSPR_{13} + RWR_{15} \times SSPR_{15} \\ &\quad + RWR_{17} \times SSPR_{17} \\ &= 0.12 \times 100 + 0.16 \times 100 + 0.12 \times 100 + 0.19 \times 100 \\ &\quad + 0.12 \times 100 + 0.12 \times 100 \\ &= 100\%\end{aligned} \tag{4.7}$$

For the four DQ dimensions in which the score was calculated, we can notice that, except for accuracy, it all has reached a good result. However, that does not represent the final score of the assessment. Just like not all defined rules have the same level of importance and not the same weight, not all DQ dimensions have the same level of importance and consequently not the same weight for calculating the final score of the DQ dimension. Having said that, to calculate the final score of the data quality in the context of this question QN1, the level of significance will be assigned for each dimension, calculating their respective weight and in the end, the final score- according to the step 8 and 9 of the methodology.

The assignment of the significance level and weight of each dimension can be found in the table 4.6. As we can see, the DQ dimension with the most expressiveness weighs is consistency. That's because it is the dimension where the rule responsible for validating the results of the COVID-19 tests is located.

Table 4.6: Level of significance and weight of each DQ dimension

DQ dimension	DQ score	Level of significance	Weight
Consistency	37.5%	10	0,4
Accuracy	100%	5	0,2
Completeness	99%	5	0,2
Duplication	100%	5	0,2

The final score for the question QN1 and up with a total of 75%. This result means that in the context of question QN1, the quality of the evaluated data was rated at 75%, which represents good quality.

$$\begin{aligned}\text{Final Score QN1} &= 0.4 \times 37.5 + 0.2 \times 100 + 0.2 \times 100 + 0.2 \times 100 \\ &= 75\%\end{aligned}$$

(2) *What is the percentage of children (patients aged up to 12 years) diagnosed with COVID-19?*

Having answered the first question, the question QN2 intended to identify the percentage of patients infected with COVID-19 that represent children. Similarly, all the steps performed and described in the question QN1 were also performed to obtain the final score result.

From the query result of question QN1, another join with the Patients and Admission table was performed to calculate the patient's age when they enter the hospital facility. A difference between the birthday of patients and the hospital admission date was performed, and all the patients whose age was between 0 and 12 were considered as children who were infected with COVID-19. Figure 4.4 demonstrates the used tables and attributes to accomplish such a result.

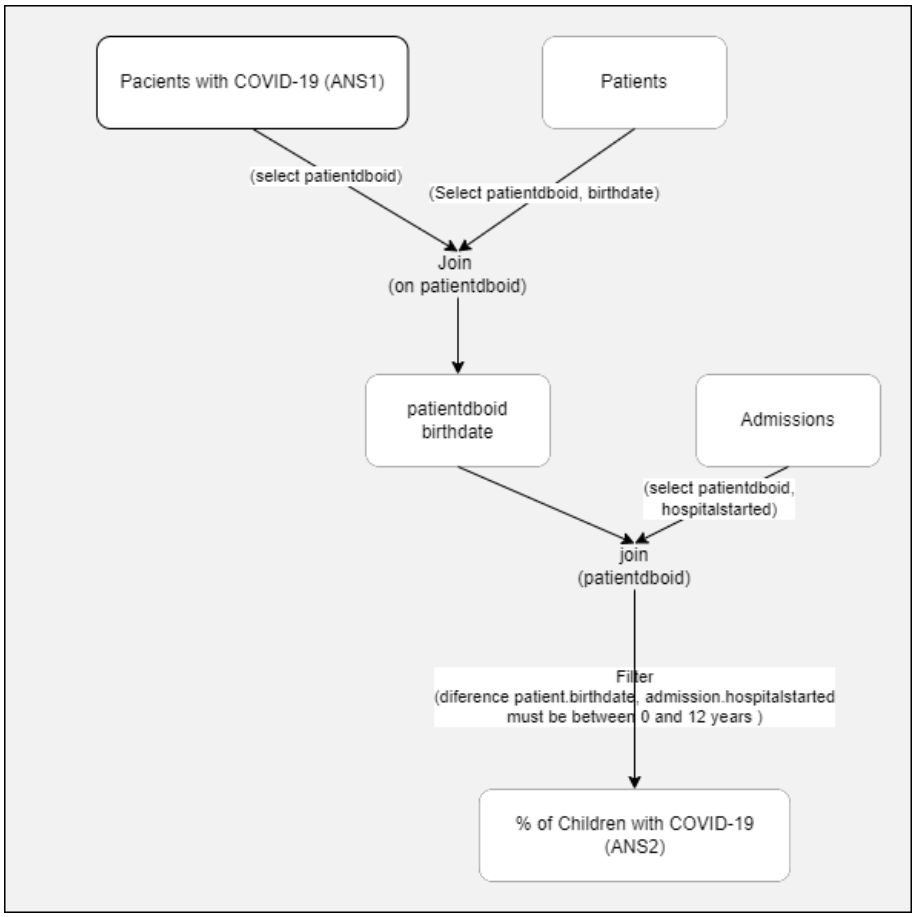


FIGURE 4.4. Tables and attributes used to determine the % of children detected with COVID-19 - QN2

Afterwards, needed rules were defined for each attribute and the corresponding DQ dimension was mapped, as shown in the table 4.7. The quality rule defined fell into two categories, namely completeness and accuracy. For each rule, "expectations" were created, allowing us to validate it and identify the percentage of success and unsuccessful of each, as listed in the table 4.8.

Table 4.7: Quality rule defined to identify % of children infected COVID-19 - QN2

Table used	Attribute name	Quality rule		Dimension
ADMISSIONS	hospitalstarted	**RL19**	hospitalstarted must not be null	Completeness
PATIENT	birthdate	**RL20**	birthdate must not be null	Completeness
		RL21	the age of the patient must be between 0 and 110	Accuracy

Table 4.8: List of defined expectations for rules of the question - QN2

Expectations	Corresponding RL	% expected	% unexpected
expect_column_values_to_not_be_null	RL19	84%	16%
expect_column_values_to_not_be_null	RL20	99%	1%
expect_column_value_lengths_to_be_between	RL21	100%	0%

To understand the validation made considering the expectations defined, Figure 4.5 is presented. In the total of three expectations defined, all of them were successfully validated regarding the data in analysis, having a score of 84% or more.

The variable age, a calculated variable resulting from the difference between the birth date of the patient and the data that the patient was admitted to the hospital, was validated with 100% of success, meaning that no patient infected with COVID-19 evaluated has the age out of the defined range, 0 - 110. Having done such analyses, it was concluded that the analysed data was successfully validated with a score of 94%, which can ensure more confidence in the presented values for the percentage of the children detected with COVID-19.

With the validation concluded, weight was calculated for both the quality rules and the dimensions mapped to them, taking into account the level of significance of each one. The results are presented in the table 4.17 and 4.10 for both rule and dimension, respectively.

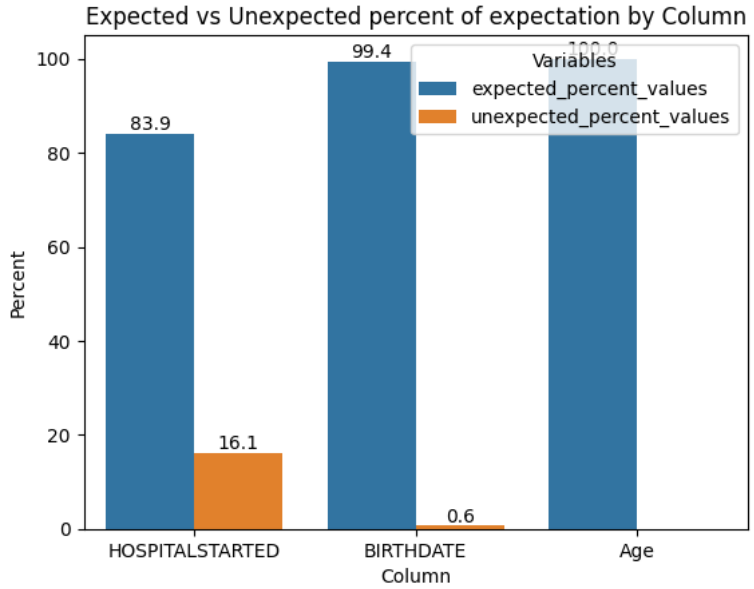


FIGURE 4.5. Expected vs Unexpected % of validated expectation by attribute - QN2

Table 4.9: Level of significance and weight of each rule of the question - QN2 - grouped by DQ dimension

Quality Rule	Dimension	Level of significance	Weight
RL19	Completeness	10	0.5
RL20	Completeness	10	0.5
RL21	Accuracy	10	1

The calculation has led to a score of 91.5% for the completeness DQ dimension and 100% for the accuracy DQ dimension.

Table 4.10: Level of significance and weight of each DQ dimension for the question - QN2

DQ dimension	DQ score	Level of significance	Weight
Accuracy	91.5%	8	0,44
Completeness	100%	10	0,56

For question QN1, the final score end up with a total of 96%.

$$\begin{aligned} \text{Final Score QN2} &= 0.44 \times 91.5 + 0.56 \times 100 \\ &= 96\% \end{aligned} \tag{4.8}$$

(3) *Does the patient present any concern about cardiac biomarkers that could lead to heart failure, specifically Troponin or (NT-pro)BNP?*

Another information that was found interesting to know is regarding the cardiac biomarkers, specifically troponin and (NT-pro)BNP, to understand if the numbers can point to concerns related to cardiac disease. Used tables and attributes to answer the proposed question are presented in 4.6. Since the patient with COVID-19 was identified, a left join was performed with the tables PICS, LABTESTS, LABRESULTS, PART-COMPOENTS and COMPONENTS. The LABTESTS table contains all the patient's laboratory tests regarding the cardiac biomarkers, and the LABRESULTS contain the results of these exams. A Filter was applied to have only laboratory tests regarding troponin and (NT-pro)BNP, and the *textvalue* attribute of the LABRESULTS table will present the test result. To evaluate these data, all the steps of the proposed methodology will be followed.

First, to calculate the data quality score regarding this question, a set of quality rules related to every attribute used was defined. The rule was categorised in their respective DQ dimension and validated using the GE by configuring expectations as listed in the table 4.11 and 4.12, respectively. A total of 10 rules were defined, falling into two DQ dimensions: Accuracy and Completeness. The configured expectation for each rule allowed for validation of the percentage of success and failure of each rule.

Table 4.11: Quality rule for question - QN3

Table used	Attribute name	Quality rule		Dimension
ADMISSION	picsdatadboid	**RL22**	picsdatadboid must belong to a valid pics record	Accuracy
LABTEST	picsdatadboid	**RL23**	picsdatadboid must belong to a pics record	Accuracy

LABRESULT	labtestdboid partcomponent-dboid textvalue	**RL24**	labtestdboid must not be null	Completeness
		RL25	labtestdboid must belong to a valid *labtest* record	Accuracy
		RL26	partcomponent-dboid must not be null	Completeness
		RL27	partcomponentd-boid must belong to a valid part compo-nent record	Accuracy
		RL28	textvalue must not be empty if the re-sult belongs to a troponin or (NT-pro)BN test	Completeness
PART-COMPONENT	componentdbo	**RL29**	componentdboid must not be null	Completeness
		RL30	componentdboid must belong to a valid *component* record	Accuracy
COMPONENT	componentdesc	**RL31**	componentdesc have at least one record related to "Troponin" or "(NT-pro)BNP	Accuracy

Table 4.12: List of defined expectations for Quality rules of the question - QN3

Expectations	Corresponding RL	% expected	% unexpected
expect_column_values_to_be_in_set	RL22	100%	0%
expect_column_values_to_not_be_null	RL23	100%	0%
expect_column_values_to_not_be_null	RL24	100%	0%
expect_column_values_to_be_in_set			

	RL25	100%	0%
expect_column_values_to_not_be_null	RL26	100%	0%
expect_column_values_to_be_in_set	RL27	100%	0%
expect_column_values_to_not_be_null (when the result belongs to troponin or (NT-pro)BN exams)	RL28	94.4%	0%
expect_column_values_to_not_be_null	RL29	100%	0%
expect_column_values_to_be_in_set	RL30	100%	0%
expect_column_values_to_match_regex	RL31	100%	0%

The figure presented in 4.7 synthesizes the validated attributes. A total of 6 attributes were validated from different tables. For all the validated rules, a 100% of success was achieved, except for the variable text value, where the success percentage achieved was 94.1%. That is because some laboratory tests regarding the cardiac biomarkers have no result registered, but the number represents a very low percentage.

After validation of each rule and having their respective percentage of success validation, the weight was calculated based on the level of significance attributed to each role, as presented in the table 4.13. The score for each dimension was also calculated.

Table 4.13: Level of significance and weight of each rule of the question - QN3 - grouped by DQ dimension

Bussiness Rule	Dimension	Level of significance	Weight
RL22	Accuracy	8	0.16
RL23	Accuracy	8	0.16
RL25	Accuracy	8	0.16
RL27	Accuracy	8	0.16
RL30	Accuracy	8	0.16
RL31	Accuracy	10	0.20
RL24	Completeness	8	0.24
RL26	Completeness	8	0.24
RL28	Completeness	10	0.29
RL29	Completeness	8	0.24

$$\begin{aligned} \text{Score Accuracy} &= 0.16 \times 100 + 0.16 \times 100 + 0.16 \times 100 + 0.16 \times 100 \\ &\quad + 0.16 \times 100 + 0.20 \times 100 \\ &= 100\% \end{aligned} \tag{4.9}$$

$$\begin{aligned} \text{Score Completeness} &= 0.24 \times 100 + 0.24 \times 100 + 0.29 \times 94 + 0.24 \times 100 \\ &= 98\% \end{aligned}$$

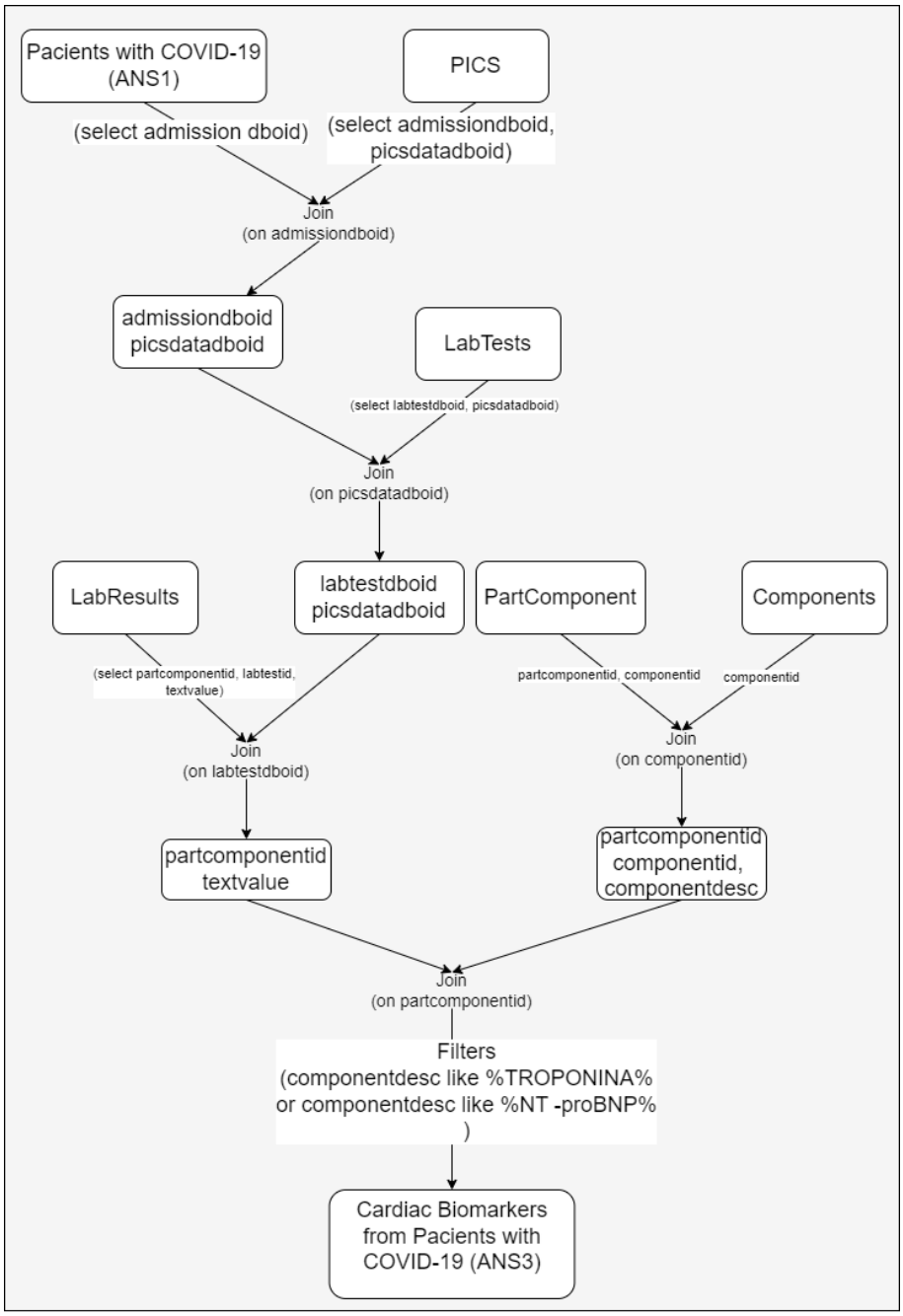


FIGURE 4.6. Tables and attributes used to identify the level of Cardiac Biomarkers

To calculate the final data quality score of each question, a level of significance was also attributed, and the respective weight was calculated as shown in the table 4.14. The score obtained was 98%.

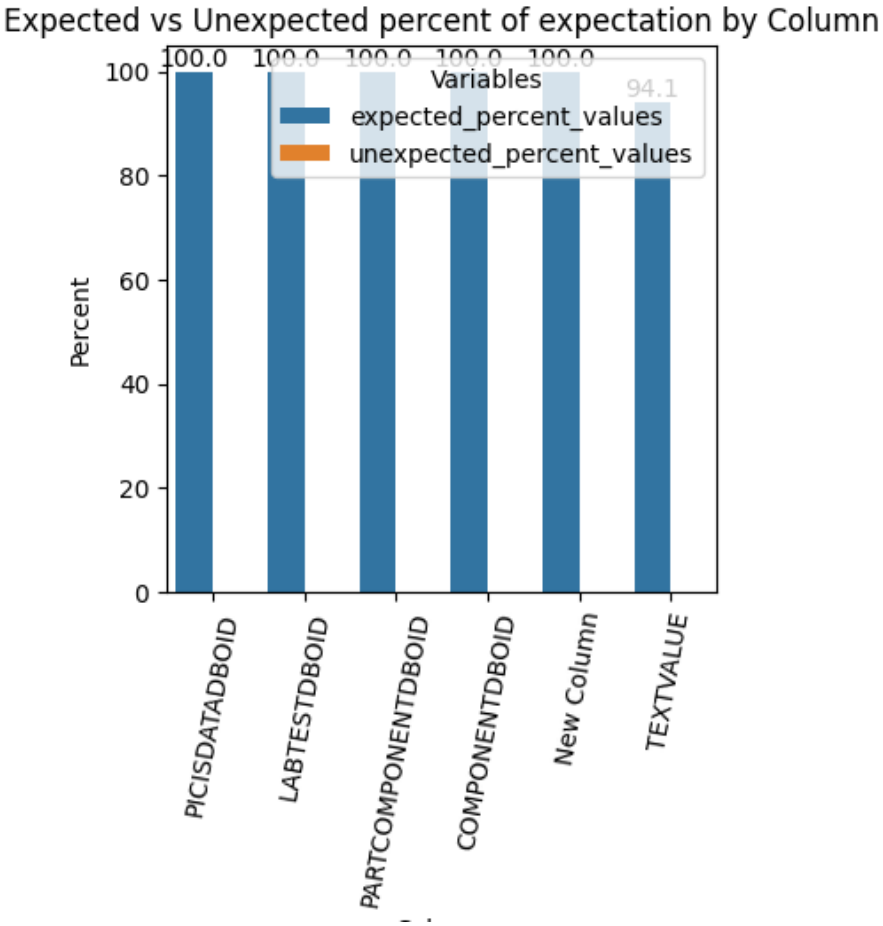


FIGURE 4.7. Expected vs Unexpected % of validated expectation by attribute - QN3

Table 4.14: Level of significance and weight of each DQ dimension for the question - QN3

Dimension	Level of success	Level of significance	Weight
Accuracy	100%	10	0.44
Completeness	98%	8	0.56

$$\begin{aligned} \text{Final Score QN4} &= 0.44 \times 100 + 0.56\times \\ &= 98.0\% \end{aligned} \tag{4.11}$$

(4) *Which ethnic group detected the highest percentage of COVID-19 infection?*

In addition, it was also interesting to know the ethnic group that was most affected by COVID-19, question QN4. The used tables and attributes to answer such questions are shown in Figure 4.8. Having the patient infected with COVID-19, a left join was performed with the Patients, Ethnicity and EtinicGroup tables. The query result was filtered only to give the patients with an ethnicity or ethnic group associated. The rest of

the process of calculating the DQ score comprises all the steps of the previously mentioned methodology.

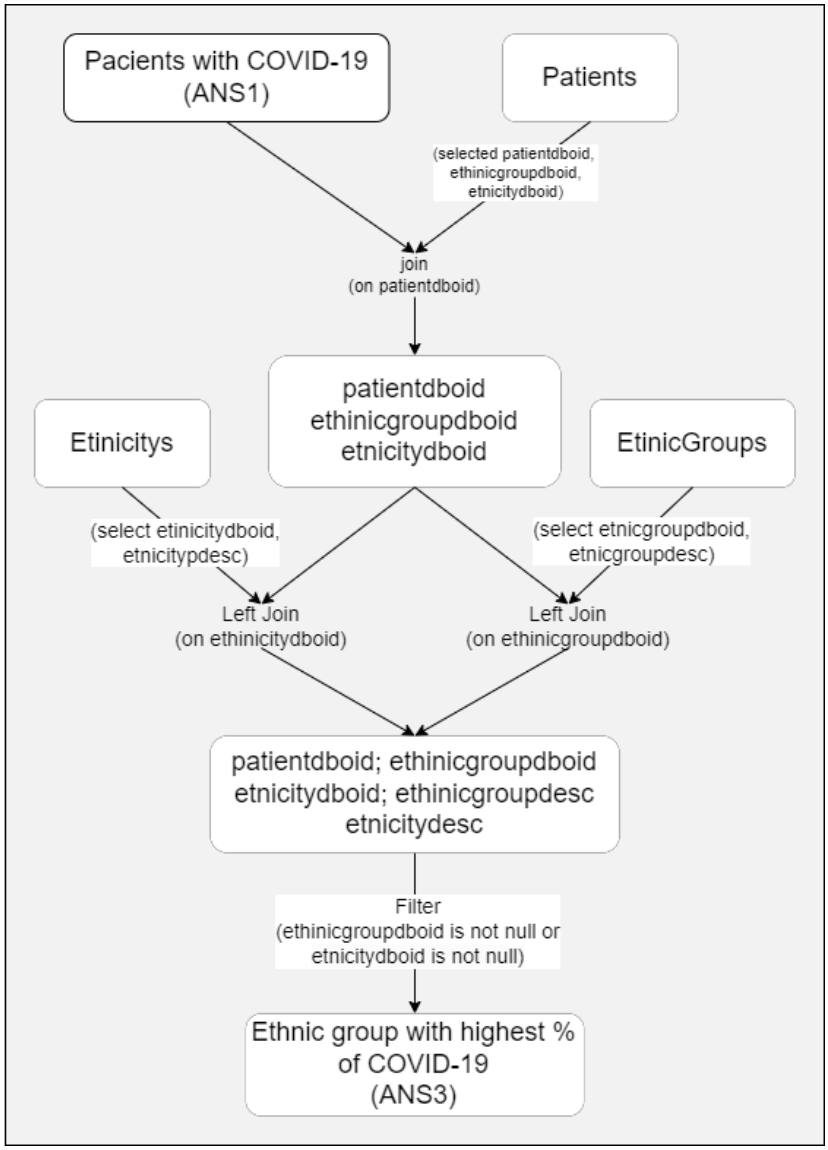


FIGURE 4.8. Tables and attributes used to % of COVID-19 detection by ethnic group - QN4

With all the necessary tables and attributes identified, the rule was defined, and for each one, the corresponding dimension was mapped according to the table 4.15. A total of 12 rule was defined and categorised into three different DQ dimension: Accuracy, Completeness and Duplication. Afterwards, a set of expectations was defined to be validated, and it is presented in the table 4.16. The validation allowed us to identify the percentage of success and failure of each one.

Table 4.15: Quality rule for question - QN4

Table used	Attribute name	Quality rule	Dimension
PATIENT	etinicgroup-dboid	**RL32** etinicgroupdboid must belong to a valid etinicgroupd	Accuracy
		RL33 etinicgroupdboid must not be null if etnicitydboid is null	Completeness
	etnicitydboid		Accuracy
		RL34 etnicitydboid must belong to a valid etnicity	Completeness
		RL35 etnicitydboid must not be null if etinicgroupdboid is null	
ETNICGROUP	ethinicgroup-dboid	**RL36** ethnicgroupdboid must not be null	Completeness
		RL37 ethnicgroupdboid must be unique	Duplication
		RL38 ethnicgroupdesc must not be duplicated	Duplication
	ethnicgroup-desc	**RL39** ethnicgroupdesc must not be empty	Completeness
ETNICITY	ethnicitydboid	**RL40** ethnicitydboid must not be null	Completeness
		RL41 ethnicitydboid must be unique	Duplication
	ethnicitydesc	**RL42** ethnicitydesc must not be duplicated	Duplication
		RL43 ethnicitydesc must not be empty	Completeness

Table 4.16: List of defined expectations for Quality rules of the question - QN4

Expectations	Corresponding RL	% expected	% unexpected
expect_column_values_to_be_in_set	RL32	0%	100%
expect_column_values_to_not_be_null (when ethnicitydboid is null)	RL33	100%	0%
expect_column_values_to_be_in_set	RL34	2%	98%
expect_column_values_to_not_be_null (when ethnicgroupdboid is null)	RL35	100%	0%
expect_column_values_to_not_be_null	RL36	100%	0%
expect_column_values_to_be_unique	RL37	100%	0%
expect_column_values_to_be_unique	RL38	100%	0%
expect_column_values_to_not_be_null	RL39	100%	0%
expect_column_values_to_not_be_null	RL40	100%	0%
expect_column_values_to_be_unique	RL41	100%	0%
expect_column_values_to_be_unique	RL42	100%	0%
expect_column_values_to_not_be_null	RL43	100%	0%

Analysing the final outcome from the validation presented in Figure 4.9, we can notice that from the four variables in the context of 3 different tables, all of them were successfully validated with a score of 75% or more. However, this general overview does not necessarily represent a positive evaluation.

Observing more details provided for each expectation validated in the table 4.16, we can see that even though the data regarding ethnic group and ethnicity present in the patient's table are always fields, in more than 98% of the cases, it is a field with data corresponding to an unknown ethnic group or ethnicity, what brings no useful information. Therefore, we can highlight that the available data is not very reliable for differentiating infected patients based on their ethnic group.

Having all the rules validated, it was grouped by dimension, as listed in the table 4.17, attributing the significance level for each one and then calculating their weight. We can notice from this question that all the rule has the same level of significance, which also has the same weight in each dimension. The score of each dimension was subsequently calculated.

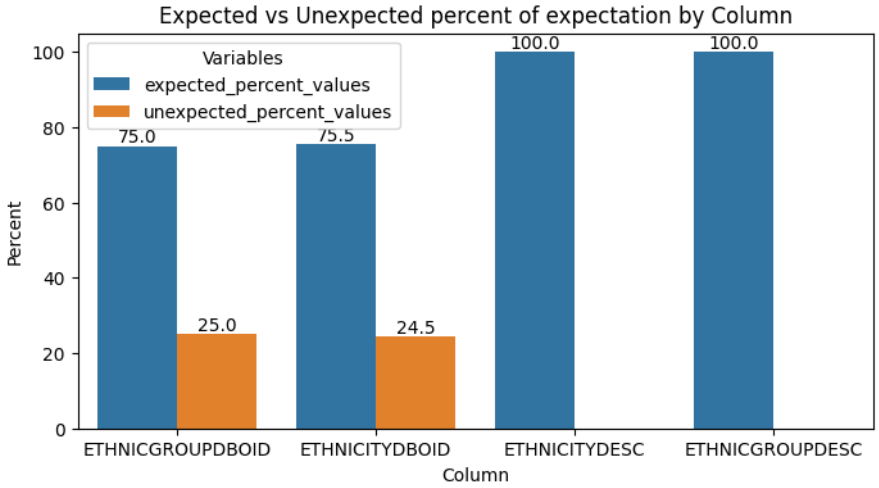


FIGURE 4.9. Expected vs Unexpected % of validated expectation by attribute - QN4

Table 4.17: Level of significance and weight of each rule of the question - QN2 - grouped by DQ dimension

Bussiness Rule	Dimenssion	Level of significance	Weight
RL32	Accuracy	10	0.50
RL34	Accuracy	10	0.50
RL33	Completeness	10	0.17
RL35	Completeness	10	0.17
RL36	Completeness	10	0.17
RL39	Completeness	10	0.17
RL40	Completeness	10	0.17
RL43	Completeness	10	0.17
RL37	Duplication	10	0.25
RL38	Duplication	10	0.25
RL41	Duplication	10	0.25
RL42	Duplication	10	0.25

$$\begin{aligned} \text{Score Accuracy} &= 0.5 \times 0 + 0.5 \times 2 \\ &= 1\% \end{aligned} \tag{4.12}$$

$$\begin{aligned} \text{Score Completeness} &= 0.17 \times 100 + 0.17 \times 100 + 0.17 \times 100 + 0.17 \times 100 \\ &\quad + 0.17 \times 100 + 0.17 \times 100 \\ &= 100\% \end{aligned} \tag{4.13}$$

$$\begin{aligned} \text{Score Duplication} &= 0.25 \times 100 + 0.25 \times 100 + 0.25 \times 100 + 0.25 \times 100 \\ &= 100\% \end{aligned} \tag{4.14}$$

To have the final score each dimension was also attributed their level of significance and respective weight as demonstrated in the table 4.18. The final DQ score calculated for this question was 23.8%.

Table 4.18: Level of significance and weight of each DQ dimension for the question - QN4

Dimension	Level of success	Level of significance	Weight
Accuracy	1%	10	0.77
Completeness	100%	2	0.15
Duplication	100%	1	0.08

$$\begin{aligned} \text{Final Score QN4} &= 0.77 \times 1 + 0.15 \times 100 + 0.08 \times 100 \\ &= 23.8\% \end{aligned} \tag{4.15}$$

(5) *What percentage of patients diagnosed with COVID-19 die?*

At last, it was found interesting to know the percentage of patients diagnosed with COVID-19 that die, question QN5. As present in Figure 4.10, from the result of the first question was performed joined with the Admission and Discharge table to have the reason why the patient was discharged. In the final, a filter was performed to have only the patient with discharge reason was "*morgue*" or "*óbito*", which represents the patient that died.

Then set of rules was also defined and listed in the table 4.19 and categorized in 3 DQ dimensions, namely Completeness, Duplication and Accuracy. The rule was evaluated taking into account the "expectations" described in the table 4.20 where the percentage of success and unsuccess of each validation can also be found.

Table 4.19: Quality rule for question - QN5

Table used	Attribute name	Quality rule		Dimension
ADMISSION	hospitalended	**RL43**	if hospitalended not null, dischdboid must not be null	Completeness
	dischdboid	**RL44**	dischdboid should belong to a valid discharge reason at least 80% of the time	Accuracy

DISCHARGE	dischdboid	**RL45**	dischdboid must not be null	Completeness
		RL46	dischdboid must be unique	Duplication
	deschdesc	**RL47**	deschdesc must not be duplicated	Duplication
		RL48	deschdesc must not be empty	Completeness

Table 4.20: List of defined expectations for rules of the question - QN5

Expectations	column	% expected	% unexpected
expect_column_values_to_not_be_null (when hospitalended is not null)	RL43	100%	0%
expect_column_values_to_be_in_set	RL44	12,9%	87,1%
expect_column_values_to_not_be_null	RL45	100%	0%
expect_column_values_to_be_unique	RL46	100%	0%
expect_column_values_to_not_be_null	RL47	100%	0%
expect_column_values_to_be_unique	RL48	100%	0%

The validated expectations lead to the results presented in Figure 4.11. Analysing it, we can notice that to answer the question QN5 two variables were validated in the context of two tables, and for both of them, only approximately 50% of the defined rule could be successfully validated. Looking into a more deep detail presented in the table 4.20 for each expectation, we can verify that the *dischdboid* that identifies the reason that the patient was discharged should be filed only when the *hospitalended* data is filled, although is always filled in that case, in 87% of the case has a record that corresponds to an unknown reason for which the patient was dismissed.

That said, only 13% of the time we can know the reason a patient was discharged and, subsequently, if the reason was death. In addition, both the *dischdboid* and *dischdesc* when analysed in the context of the discharge table, we verify that they are not unique because 50% of the data presented in that table was duplicated. These facts lead us to highlight that, according to analysed data, a low number of patients is known if the discharge reason was because of death. However, the absence of such information doesn't mean these data don't exist. Since often relevant information is kipped in clinical diaries, which are non-structured data, more advanced mechanisms are required to extract this information, which naturally affects the time and effort needed to solve it.

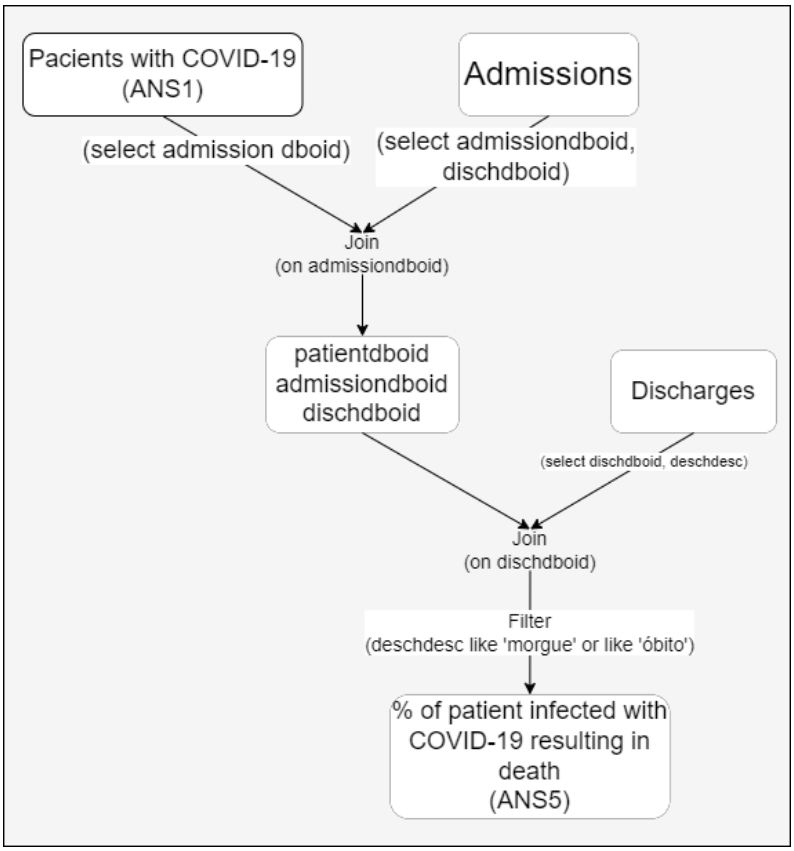


FIGURE 4.10. Used tables and attributes to answer the question - QN5

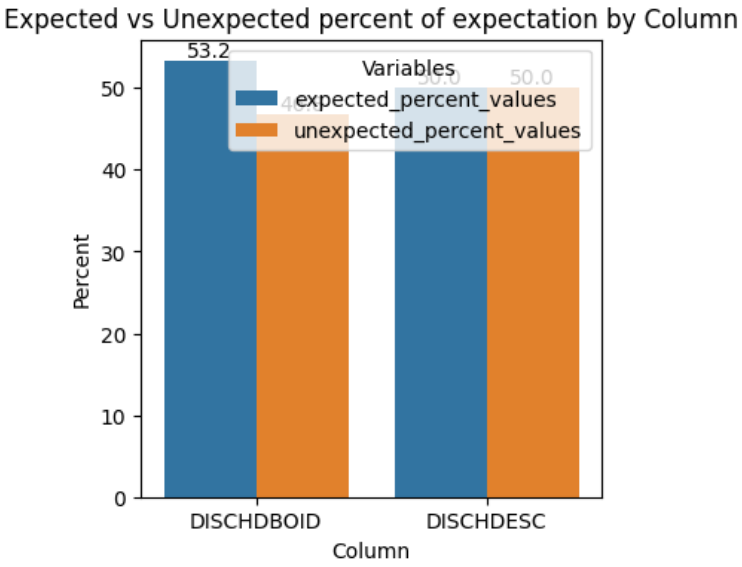


FIGURE 4.11. Expected vs Unexpected % of validated expectation by attribute - QN4

With the validation of the defined rules completed, they were grouped by DQ dimension, and the level of significance of each one was attributed and used to calculate their respective weight as presented in the table 4.21. The score of each dimension was later calculated. For both the Completeness and Duplication have a score of 100% meaning

that all the rule defined for these dimensions was successfully validated, however for the accuracy dimension the score was only 13%.

Table 4.21: Level of significance and weight of each rule of the question -QN5 - grouped by DQ dimension

Quality Rule	Dimension	Level of significance	Weight
RL44	Accuracy	10	1.00
RL43	Completeness	6	0.33
RL45	Completeness	6	0.33
RL48	Completeness	6	0.33
RL46	Duplication	3	0.50
RL47	Duplication	3	0.50

$$\begin{aligned} \text{Score Accuracy} &= 1 \times 13 \\ &= 13\% \end{aligned} \tag{4.16}$$

$$\begin{aligned} \text{Score Completeness} &= 0.33 \times 100 + 0.33 \times 100 + 0.33\times \\ &= 100\% \end{aligned} \tag{4.17}$$

$$\begin{aligned} \text{Score Duplication} &= 0.5 \times 100 + 0.5 \times 100 \\ &= 100\% \end{aligned} \tag{4.18}$$

To have the final DQ score for this question, the dimension was also their respective level of significance, and the weight was calculated as shown in the table 4.22. The final DQ score calculated was 57.0%.

Table 4.22: Level of significance and weight of each DQ dimension forthe question - QN5

Dimension	Level of success	Level of significance	Weight
Accuracy	13%	10	0.5
Completeness	100%	2	0.25
Duplication	100%	5	0.25

$$\begin{aligned} \text{Final Score QN5} &= 0.5 \times 13 + 0.25 \times 100 + 0.25 \times 100 \\ &= 57.0\% \end{aligned} \tag{4.19}$$

4.4. Evaluation and results

This section aims to present the results regarding the DQ score of the five questions proposed in the context of the data validation. The obtained result is presented in figure 4.13, and a brief interpretation will be conducted to better understand it.

Analysing Figure 4.13 can notice that to answer the proposed questions, a total of 8 tables and 27 distinct attributes were used. For these attributes, 48 rules were created and mapped to the corresponding DQ dimensions. The mapping of the rules with the DQ dimensions resulted in them being categorised into four distinct dimensions, namely accuracy, completeness, duplication and consistency, where 46% of rules fell into completeness, 25% into accuracy, 27% in duplication and 2% in consistency.

Going deep into each question, it was noted that, regarding question QN1, which aims to identify the patient detected with COVID-19, a total of 18 rules were categorised, where eight were categorised as completeness, 7 were duplication, 2 were accuracy and 1 were consistency. Regarding the weight associated with each dimension, we notice that the dimension consistency has the highest weight, with 40% in 100%, even though it is associated with a single rule. This is justified due to the fact that the rule with which it is associated is crucial to determine if the patient has COVID-19 or not since it is related to the result of the COVID-19 test performed. The other three dimensions, accuracy, duplication and completeness, are shown as having equal weight, meaning they have the same level of importance. Even though most of the rule was mapped to completeness and duplication DQ dimensions, they do not represent the most significant ones, meaning that the rule associated with these dimensions has less relevance than the one associated with the consistency dimension.

The final score obtained regarding the DQ of the source database when evaluated in the context of question QN1 was 0.75, meaning that the data made available can guarantee that we can identify the patient with COVID-19 with 75% confidence. This is due to a lack of consistency when recording the laboratory results of tests carried out to detect COVID-19 in patients.

Taking into account the question QN2, in which the goal is to identify the percentage of children with COVID-19, we can observe that to answer that question, only three rules were created, two of which fell into the completeness dimension and one into the accuracy dimension. The completeness dimension was attributed a total weight of 0,56 and an accuracy of 0,44. Calculating the score regarding this question it was obtained a final result of 96%, which means that for the patients identified with COVID-19, we can affirm with 96% confidence if they are children or not. That's because the age of the patients was known by the majority of the record.

About the question QN3, which aims to know if the patients have levels of the cardiac biomarkers that could point to heart failure, a total of ten rules was defined and included in 2 DQ dimensions, Accuracy and Completeness. The accuracy dimension has attributed a weight of 0.44 while the completeness 0.56 since it's is essential to have all the laboratory test results registered to answer the proposed question. By calculating the DQ score of this question, the result obtained was 98%, representing that the data analysed to answer this question have an outstanding quality. Therefore, the answer to the question could be given with an excellent level of confidence.

Concerning question QN4, which pretends to categorise the patients infected with COVID-19 by ethnicity, we can verify that a total of 12 rule was analysed, where six was included in the dimension completeness, four dimension duplication and two dimensions accuracy. The DQ dimension with the highest weight was accuracy, having 0,77 of the total weight, followed by completeness with 0,15 and duplication with 0,08. The final score obtained concerning the data quality in the context of such a question was 0.24, meaning that we have only 24% confidence when differentiating patients diagnosed with COVID-19 by ethnicity. This low confidence level regarding such specifications is because a very low percentage of these patients have specified their ethnicity. Most of them have their ethnicity classified as "others", which gives no meaningful information.

Lastly, about question QN5, which intends to identify the patient with COVID-19 that ended in death, a total of 6 rules were created, 3 of which were included in completeness DQ dimensions, 2 in duplication and 1 in accuracy. The accuracy dimension has attributed a weight of 0.5 despite having only one rule associated. That's because the rule associated is responsible for validating the reason the patient was discharged from the hospital, whether it is death or not. This is crucial to answer the proposed question. The other 0.5 is attributed to completeness and duplication, rated with 0.25 each. The calculation of the final score, regarding the quality of the available data to answer the question in context, was 0.57, meaning that we can only identify, in useful time, if the patient infected with COVID-19 died with 57% confidence because in a lot of cases where these patients were discharged, was not indicate the reason of the same. However, this result does not mean that this information does not exist. Since much information regarding the patients is kept in clinic notes, this more advanced mechanism would be necessary to verify the existence of this data.

This analysis leads us to conclude that of the five questions proposed, only two achieved a score of more than 80%. When the main goal of the Capacity project is to migrate data from patients with COVID-19, not being able to identify this patient with more than 80% confidence can be alert to put in place measures that can improve the quality of the available and future data from the perspective of the different dimensions evaluated.

Figure 4.12 aims to emphasise how the analysis made based on the proposed question can compromise populating the Capacity database tables.

Regarding the first question, which pretends to identify patients with COVID-19, the data analysed impacts the "Inclusion Criteria" table of the Capacity database since the prerequisite for the patient to be included in the study is to have been infected with COVID-19. The second and fourth question intends to identify the children diagnosed with COVID-19 and differentiate these patients by ethnicity. They are directly related to the table "Demographic" table of the Capacity database since it is where these data would be stored. For this table, special attention should be taken once the analysed subset of data resulted in a deficient score regarding the data quality of the second question. This may be due to the case that the majority of the patient have their ethnicity registered as

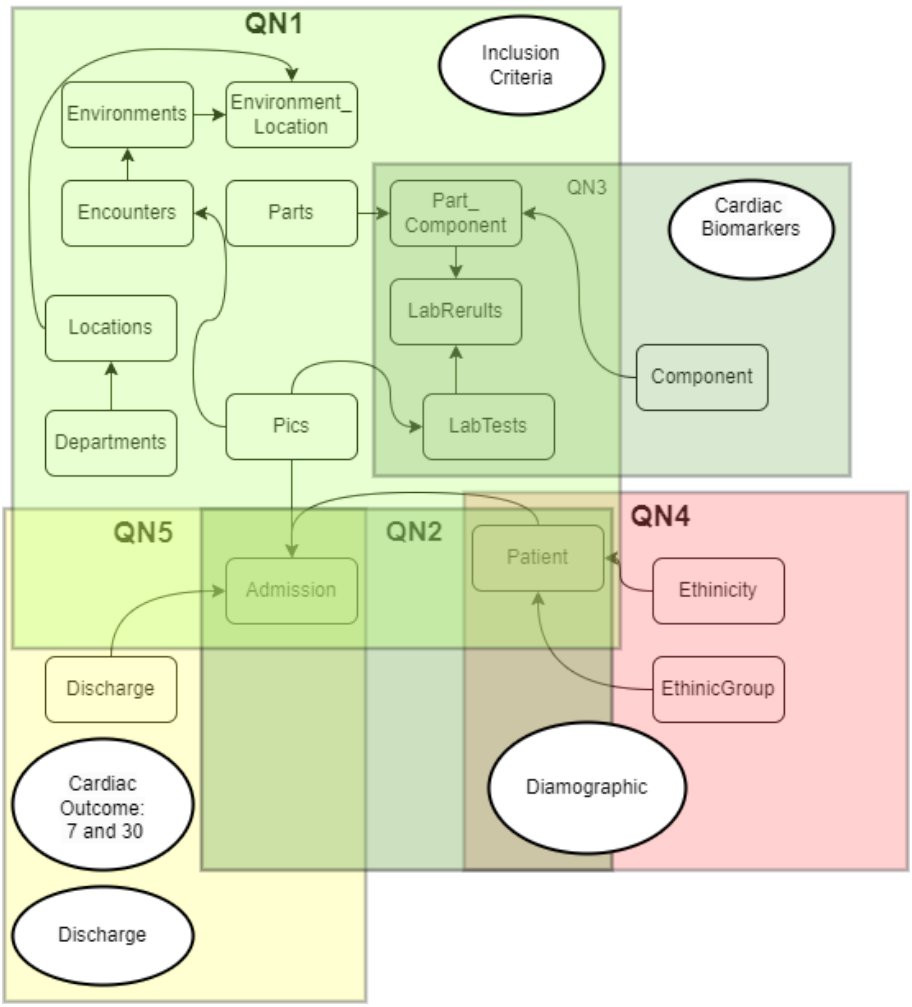


FIGURE 4.12. Relationship between the proposed questions and the Capacity tables affected

unknown. Actions should be taken to address such a problem. In concern with question 3, which intends to know the cardiac biomarkers of the patient diagnosed with COVID-19, this directly affects the table "Cardiac Biomarkers" of the Capacity database. Nevertheless, based on the analyses, no concern has been raised since the data showed an excellent quality score. Lastly, regarding question five, which intends to identify the reason for the death of the patient diagnosed with COVID-19, the analysis directly impacts the tables "Cardiac Outcome 7 days", Cardiac Outcome 30 days" and the table "Discharge" once all of this table save information regarding the death of the patients. Some attention is also recommended for these data since the data quality score is around 57%.

FIGURE 4.13. Proposed analytical question dashboard

CHAPTER 5

Conclusions

The concluding chapter of this thesis presents a comprehensive summary regarding the main findings and implications of the study on data quality in health care. Through a deep analysis of the essential literature, this academic work highlights the crucial importance of data quality in the data integration process, essentially for scenarios of the data-driven decision, as well as the proposed methodologies to implement a process of data quality evaluation. The main result of the application of some methodology appointed in the literature review, in the practical scenario of the evaluation of a subset data of the Santa Maria Hospital, are summarised, and direction for future work are also pointed out, opening opportunities for new possibilities and improvements in the evaluation of data quality in the healthcare field.

5.1. Discussion

First, the related work was crucial in understanding the important contributions of the various studies in the data quality fields. When we mention decision-making, data is the main driving force behind it in today's scenarios. Regardless of the activity area, data-driven decision-making has proven its efficiency and the positive impact it can have on an organisation. In healthcare, this is also a fact, supporting direct patient care and coordination among healthcare providers [**21**]. The recent COVID-19 pandemic reaffirms the importance of reliable data. However, in this era, where a massive amount of data is produced daily, low-quality data will likely be produced, and therefore, particular attention regarding data quality should be paid. In health care data, where decision-making often represents the difference between the life and death of patients or even the mass population, this attention must be redoubled.

If it's true the importance of high-quality data in the healthcare field [**21**], it's also true that it is particularly challenging to achieve such a feat in this area, much of it due to the complexity of the area in question and the heterogeneity of the EHR in the market, especially when we talking about data integration. Therefore, the existing literature has proposed some methodology to facilitate this hard work. The review literature also showed that the data quality evaluation is essential to ensure the data quality, and the data quality dimension is indispensable to provide such evaluation [**27**]. Numerous data quality dimensions can be evaluated in the context of data quality assessment. Still, according to [**6**], in healthcare, particularly 7, data quality dimensions have been proposed: Accuracy, Completeness, Consistency, Currency, Usability, Relevance and Duplication. Evaluating such dimension can be critical, but [**27**] appointed that, when added to quality rules, the

evaluation can be even more relevant since the concept of data quality is not universal and may vary depending on the application domain. To ensure data quality, a set of quality rules needs to be established, which includes constraints on data generation, entry, and creation. These rules can be created or discovered to correct or eliminate this data's poor data quality. Another methodology proposed the addition of weight to the evaluation process of the data quality since [**29**] defended that different data can have different impacts on the business purpose. Since some data can be more valuable, it should greatly impact the evaluation of data quality. Knowing the different points of view of different authors, for this academic work, a methodology to evaluate the quality of the data was proposed, evaluating the data by taking into account the rule, the weight and the data quality dimensions.

A practical case study was implemented to apply all knowledge retained. The main goal of this case study was to evaluate the quality of a subset of data from Santa Maria Hospital in the context of the population of a European database, Capacity, which is a project that aims to collect data on patients diagnosed with COVID-19 from different countries across Europe, to study the correlation with cardiovascular diseases. The first step for that evaluation was analysing and understanding the source database. It was composed of 138 tables with distinct scopes and particularities. However, not all of these tables were analysed. A mapping between the Hospital Santa Maria and the Capacity database was conducted, identifying the main tables and attributes necessary to populate the target database, resulting in the selection of 21 tables. The analysis of these tables reveals relevant information regarding patient admission, patient data and demographic information, laboratory test results, psychological data, diseases and treatments and so on. Through this analysis, it was possible to achieve the first objective set for this thesis, identifying patients with COVID-19 to determine if they are eligible for the Capacity study.

After analysing the selected tables and understanding the main data provided, a data profile analysis occurred, according to the second object defined using *ydata-profiling tool*. The conducted analysis resulted in a deep understanding of each variable, knowing its purpose, the type of data held, and the percentage of null and duplication values. An in-depth analysis allowed us to identify the main issues regarding each attribute and do a corresponding mapping with each data quality dimension violated. Issues such as domain violation, heterogeneity of unit measures, presence of outliers, huge percentage of null values, primary key violation, duplicate rows, misspelling error, incoherent date and so on were identified and mapped for distinct data quality dimensions, such as *Completeness*, *Accuracy*, *Duplication* and *Consistency*. Considering the main issues identified, a set of rules was defined to help appropriately address these issues and avoid future ones, reaching the third objective defined.

This dissertation's last step was to assess the quality of a subset of data from Santa Maria Hospital data, proposed in the fourth objective defined. Five analytical questions

were proposed, and an evaluation was made for each. These questions were considered critical and allowed the data quality analysis without using irrelevant columns/tables having a positive or negative impact on the overall assessment. The assessment was based on a proposed methodology that compressed nine essential steps. The first step of the methodology was to identify the question itself. After the necessary tables and attributes to answer each question. For that, joins between tables were performed, and the main attributes were selected. Afterwards, the necessary rules to ensure the quality for each attribute were defined and mapped with the corresponding DQ dimension. Ensuring each rule's success would ensure the data quality in the context of each question. The defined rules were subsequently validated using the GE framework, indicating each validation's success percentage. The rules were also grouped by their corresponding data DQ dimension and attributed a significance level on a scale of 1 to 10, where 1 means less importance and 10 means high importance. The corresponding weight for each attribute was calculated through the significance level. The sum of the weight for dimensions should be 1. After, the score of each dimension was calculated, namely the dimension Accuracy, Completeness, Duplication and Consistency depending on each one was used in each question. However, the score of each dimension was essential to know, the final goal was the DQ score of the question. For that, the same process was applied to the dimension level, attributing the significance level for each dimension and calculating the weight and the final DQ score of the question.

The overall result of the performed assessment was satisfactory. Although we have questions reaching a score of 98%, others are reaching only 24%. When we mention question QN1, essential to identify the patient with COVID-19, a DQ score of 75% has room for improvement. To guarantee such improvement, it is essential to standardise the laboratory test results for COVID-19 since they were the leading cause that negatively contributed to lowering the score. About the question QN4, which allows us to understand the ethnic group most affected by COVID-19, the score was very low. This is due to the reason that kind of information regarding the patient is not kept in most of the cases. Action to ensure the obligation of introducing these data in the systems could be a possible solution to minimise such lack of information. Regarding the question QN5, which allows us to see the percentage of patients infected with COVID-19 that died, the DQ score did not exceed 57%. That is because, in many cases, why a patient was discharged is not provided in the analysed data. However, this doesn't mean that information doesn't exist. As previously mentioned, much information regarding patients is kept in a text-free note. Using NPL to extract this information could be a path to improve the data quality score of this question. Regarding the question QN2 and QN3, an excellent data quality score was achieved.

The calculation of the DQ score for each question allowed us to achieve the fifth objective and last objective defined for this thesis.

5.2. Future Work

The development of this work has brought some contributions to the overall scenario of the data quality assessment in health care. But it also opens the way for future contributions and opportunities.

One promising direction for future work is implementing *Natural Language Processing* as part of the proposed methodology to extract meaningful information from non-structured data and its subsequent evaluation. Since in the healthcare field, the presence of non-structure data is highly probable due to the number of free medics' notes daily, these data are often hard to evaluate. In the case study addressed in this thesis, a considerable amount of the data needed to populate the Capacity database was available in these clinical notes that were not extracted and subsequently not evaluated. Future implementation of the NLP layer will enrich the assessment, making it more robust.

Another area for future exploration would be including machine learning algorithms in the data quality evaluation process. Some studies have been conducted in this field, and exciting results trying to detect problems regarding data quality dimensions have been presented, such as completeness and accuracy [**33**]. Due to the capacity of machine learning to work with high amounts of data, exploring this subject in the healthcare context can bring significant insights to the literature.

Since a very important part of this academic work relies on validating quality rules through the Great Expectation Framework, automating the rule validation process would also be an interesting field to explore. In the healthcare field, new data are constantly collected, and new rules to ensure their quality are also necessary. Having the process automated would make data evaluation more efficient and could contribute to significant improvements in terms of data quality.

Lastly, expanding the data source can also be an exciting direction for future work. Expanding the data quality analysis for other Portuguese medical facilities can be challenging but promising to understand the overall panorama of the data quality in healthcare at the national level and take joint action to address prominent issues and improve such an important area that can positively impact national health.

In conclusion, this thesis highlights the paramount importance of data quality in healthcare. All the initially proposed objectives were successfully achieved. The outcomes reaffirm the importance of data quality in an era where data is easily generated and obtained. Although the amount of information is significant for decision-makers, what brings advantages to the table is the quality of this information and how reliable it is. However, this is not always an easy task. In complex domains, such as healthcare, the implementation of a robust methodology for data quality assessment can be the ideal path to achieve such accomplishment. Constant evaluation and monitoring can also be an ally in such a process.

Bibliography

[1] M. J. Page, J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, J. M. Tetzlaff, E. A. Akl, S. E. Brennan, *et al.*, "The prisma 2020 statement: an updated guideline for reporting systematic reviews," *International journal of surgery*, vol. 88, p. 105906, 2021.

[2] D. Faggella, "Where healthcare's big data actually comes from," Nov 2019.

[3] M.-T. Chen and T.-H. Lin, "A provable and secure patient electronic health record fair exchange scheme for health information systems," *Applied Sciences*, vol. 11, no. 5, p. 2401, 2021.

[4] F. Khennou, N. E. H. Chaoui, and Y. I. Khamlichi, "A migration methodology from legacy to new electronic health record based openehr," *International Journal of E-Health and Medical Communications (IJEHMC)*, vol. 10, no. 1, pp. 55–75, 2019.

[5] Y. Li, J. Yang, Z. Zhang, J. Wen, and P. Kumar, "Healthcare data quality assessment for cybersecurity intelligence," *IEEE Transactions on Industrial Informatics*, vol. 19, no. 1, pp. 841–848, 2022.

[6] C. Liu, A. Talaei-Khoei, V. C. Storey, and G. Peng, "A review of the state of the art of data quality in healthcare.," *Journal of Global Information Management*, vol. 31, no. 1, 2023.

[7] L. A. Kapsner, J. M. Mang, S. Mate, S. A. Seuchter, A. Vengadeswaran, F. Bathelt, N. Deppenwiese, D. Kadioglu, D. Kraska, and H.-U. Prokosch, "Linking a consortium-wide data quality assessment tool with the miracum metadata repository," *Applied Clinical Informatics*, vol. 12, no. 04, pp. 826–835, 2021.

[8] K. Honeyford, P. Expert, E. Mendelsohn, B. Post, A. Faisal, B. Glampson, E. Mayer, and C. Costelloe, "Challenges and recommendations for high quality research using electronic health records," *Frontiers in Digital Health*, vol. 4, p. 940330, 2022.

[9] "Capacity registry." https://capacity-covid.eu/. [Accessed 20-06-2023].

[10] J. Cunha, R. Duarte, T. Guimarães, and M. F. Santos, "Openehr and business intelligence in healthcare: an overview," *Procedia Computer Science*, vol. 220, pp. 874–879, 2023.

[11] K. Kaloyanova, I. Naydenova, and Z. Kovacheva, "Addressing quality issues in secondary use of health data," 2022.

[12] J. A. Aunger, R. Millar, A. M. Rafferty, and R. Mannion, "Collaboration over competition? regulatory reform and inter-organisational relations in the nhs amidst the covid-19 pandemic: a qualitative study," *BMC Health Services Research*, vol. 22, no. 1, pp. 1–15, 2022.

[13] M.-T. Chen and T.-H. Lin, "A provable and secure patient electronic health record fair exchange scheme for health information systems," *Applied Sciences*, vol. 11, no. 5, p. 2401, 2021.

[14] H. N. Benkhaled and D. Berrabah, "Data quality management for data warehouse systems: State of thes art.," *JERI*, 2019.

[15] T. Z. Ali, T. M. Abdelaziz, A. M. Maatuk, and S. M. Elakeili, "A framework for improving data quality in data warehouse: A case study," in *2020 21st International Arab Conference on Information Technology (ACIT)*, pp. 1–8, IEEE, 2020.

[16] M. Souibgui, F. Atigui, S. Zammali, S. Cherfi, and S. B. Yahia, "Data quality in etl process: A preliminary study," *Procedia Computer Science*, vol. 159, pp. 676–687, 2019.

[17] S. Moorthie, S. Hayat, Y. Zhang, K. Parkin, V. Philips, A. Bale, R. Duschinsky, T. Ford, and A. Moore, "Rapid systematic review to identify key barriers to access, linkage, and use of local

authority administrative data for population health research, practice, and policy in the united kingdom," *BMC Public Health*, vol. 22, no. 1, pp. 1–13, 2022.

[18] P. L. Mai, S. R. Sand, N. Saha, M. Oberti, T. Dolafi, L. DiGianni, E. J. Root, X. Kong, R. C. Bremer, K. M. Santiago, *et al.*, "Li-fraumeni exploration consortium data coordinating center: building an interactive web-based resource for collaborative international cancer epidemiology research for a rare condition," *Cancer Epidemiology, Biomarkers & Prevention*, vol. 29, no. 5, pp. 927–935, 2020.

[19] N. Oubenali, S. Messaoud, A. Filiot, A. Lamer, and P. Andrey, "Visualization of medical concepts represented using word embeddings: a scoping review," *BMC Medical Informatics and Decision Making*, vol. 22, no. 1, pp. 1–14, 2022.

[20] T. Zhang, A. M. Schoene, S. Ji, and S. Ananiadou, "Natural language processing applied to mental illness detection: a narrative review," *NPJ digital medicine*, vol. 5, no. 1, p. 46, 2022.

[21] C. Batini, M. Scannapieco, *et al.*, "Data and information quality," *Cham, Switzerland: Springer International Publishing*, 2016.

[22] L. Ehrlinger and W. Wöß, "A survey of data quality measurement and monitoring tools," *Frontiers in big data*, p. 28, 2022.

[23] S. Juddoo, C. George, P. Duquenoy, and D. Windridge, "Data governance in the health industry: Investigating data quality dimensions within a big data context," *Applied System Innovation*, vol. 1, no. 4, p. 43, 2018.

[24] Á. Valencia-Parra, L. Parody, Á. J. Varela-Vaca, I. Caballero, and M. T. Gómez-López, "Dmn for data quality measurement and assessment," in *Business Process Management Workshops: BPM 2019 International Workshops, Vienna, Austria, September 1–6, 2019, Revised Selected Papers 17*, pp. 362–374, Springer, 2019.

[25] F. Hardy, J. Heyl, K. Tucker, A. Hopper, M. J. Marchã, T. W. Briggs, J. Yates, J. Day, A. Wheeler, S. Eve-Jones, *et al.*, "Data consistency in the english hospital episodes statistics database," *BMJ Health & Care Informatics*, vol. 29, no. 1, 2022.

[26] A. Ngueilbaye, J. Z. Huang, M. Khan, and H. Wang, "Data quality model for assessing public covid-19 big datasets," *The Journal of Supercomputing*, pp. 1–33, 2023.

[27] I. Taleb, M. Serhani, C. Bouhaddioui, and R. Dssouli, "Big data quality framework: a holistic approach to continuous quality management. j. big data 8 (1), 1–41 (2021)."

[28] N. S. Rajan, R. Gouripeddi, P. Mo, R. K. Madsen, and J. C. Facelli, "Towards a content agnostic computable knowledge repository for data quality assessment," *Computer methods and programs in biomedicine*, vol. 177, pp. 193–201, 2019.

[29] W. Elouataoui, I. El Alaoui, S. El Mendili, and Y. Gahi, "An advanced big data quality framework based on weighted metrics," *Big Data and Cognitive Computing*, vol. 6, no. 4, p. 153, 2022.

[30] R. Kimball and J. Caserta, *The data warehouse ETL toolkit*. John Wiley & Sons, 2004.

[31] "pandas - Python Data Analysis Library — pandas.pydata.org." https://docs.profiling.ydata.ai/4.6/. [Accessed 25-10-2023].

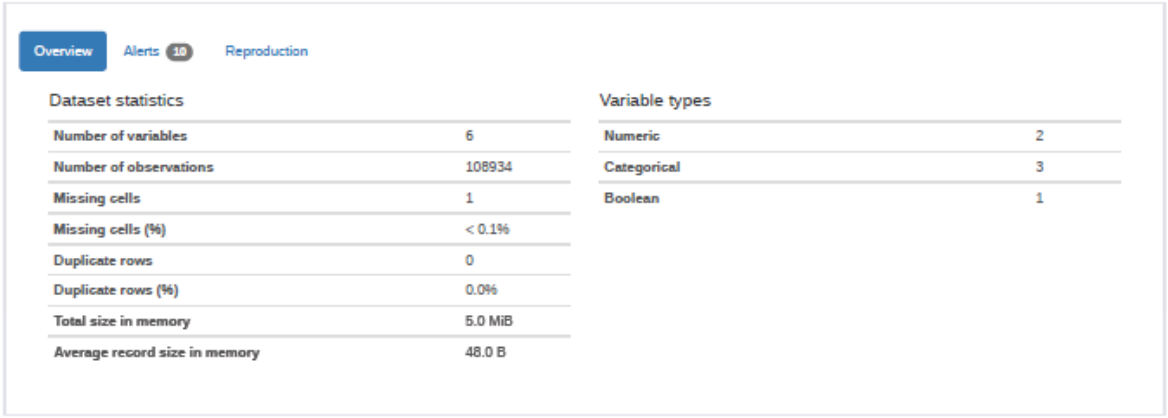
[32] "Great expectations." https://docs.greatexpectations.io/docs/home/. [Accessed 18-08-2023].

[33] S. Juddoo and C. George, "A qualitative assessment of machine learning support for detecting data completeness and accuracy issues to improve data analytics in big data for the healthcare industry," in *2020 3rd International Conference on Emerging Trends in Electrical, Electronic and Communications Engineering (ELECOM)*, pp. 58–66, IEEE, 2020.

APPENDIX A

Discharge table data profiling report

Overview



Overview | Alerts 19 | Reproduction

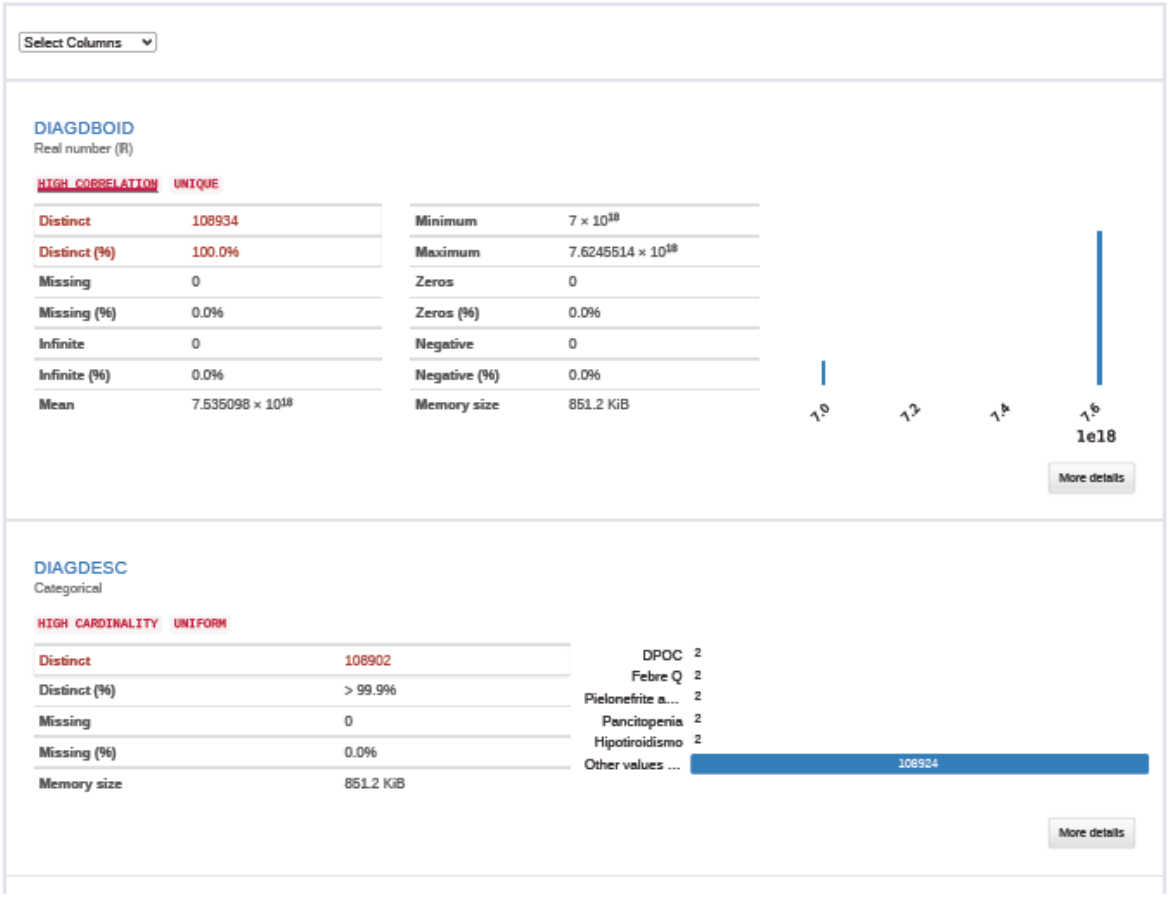
Dataset statistics

Number of variables	6
Number of observations	108934
Missing cells	1
Missing cells (%)	< 0.1%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	5.0 MiB
Average record size in memory	48.0 B

Variable types

Numeric	2
Categorical	3
Boolean	1

Variables



Select Columns

DIAGDBOID

Real number (R)

HIGH CORRELATION UNIQUE

Distinct	108934	Minimum	7×10^{18}
Distinct (%)	100.0%	Maximum	7.6245514×10^{18}
Missing	0	Zeros	0
Missing (%)	0.0%	Zeros (%)	0.0%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	7.535098×10^{18}	Memory size	851.2 KiB

More details

DIAGDESC

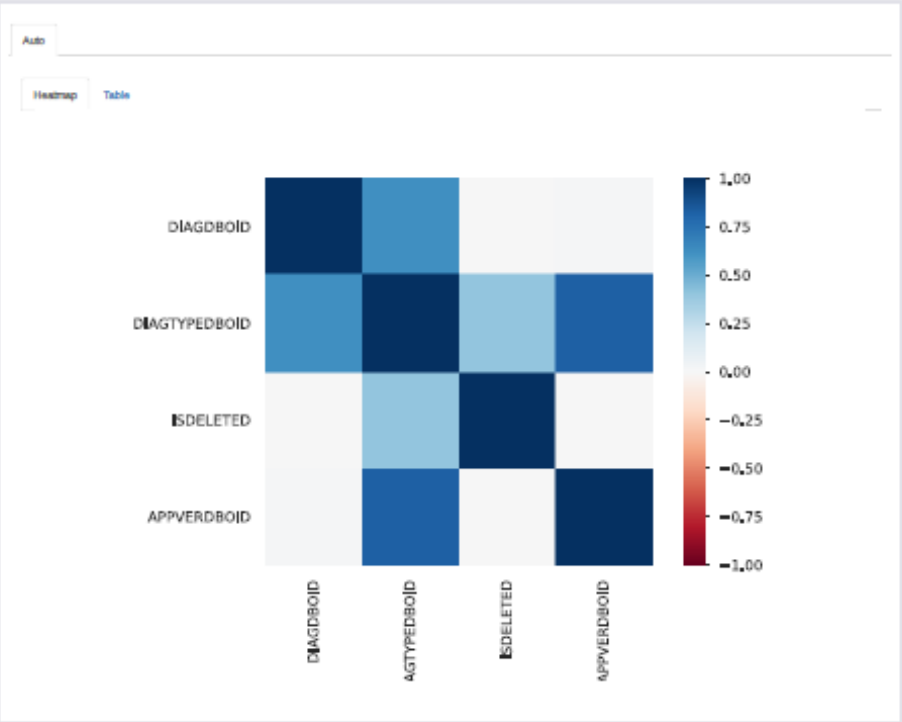
Categorical

HIGH CARDINALITY UNIFORM

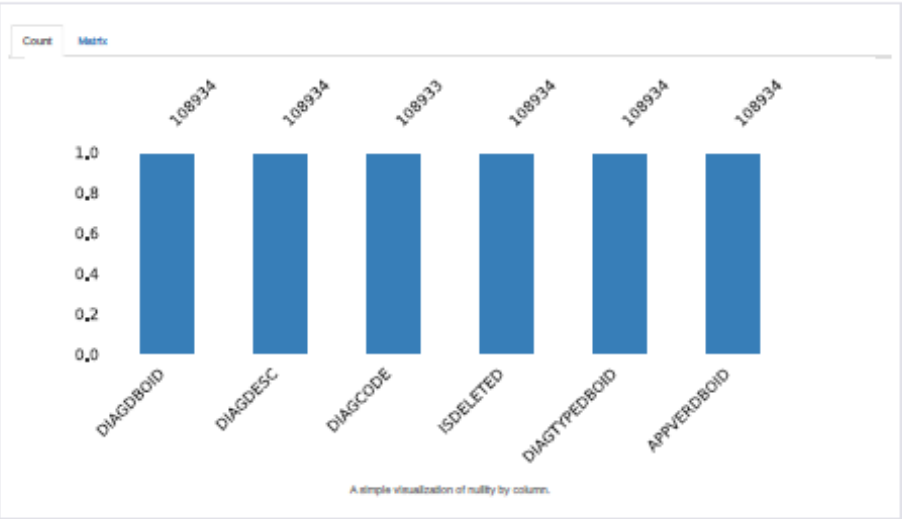
Distinct	108902
Distinct (%)	> 99.9%
Missing	0
Missing (%)	0.0%
Memory size	851.2 KiB

More details

Correlations



Missing values



Sample

First rows | Last rows

	DIAGDBOID	DIAGDESC	DIAGCODE	ISDELETED	DIAGTYPEDBOID	APPVERDBOID
0	70000000097560000552	INFECCOES ESPECIFICAS DO PERIODO PERINATAL	771	F	60000000000000000552	27000000000001600
1	70000000097570000552	RUBEOLA CONGENITA NO PERIODO PERINATAL	7710	F	60000000000000000552	27000000000001600
2	70000000097580000552	INFECCAO CONGENITA POR CITOMEGALOVIRUS	7711	F	60000000000000000552	27000000000001600
3	70000000097590000552	INFECCOES CONGENITAS NCOP	7712	F	60000000000000000552	27000000000001600
4	70000000097600000552	TETANO NEONATAL	7713	F	60000000000000000552	27000000000001600
5	70000000097610000552	ONFALITE DO RECEM-NASCIDO	7714	F	60000000000000000552	27000000000001600
6	70000000097620000552	MASTITE NEONATAL INFECCIOSA	7715	F	60000000000000000552	27000000000001600
7	70000000097630000552	CONJUNTIVITE E DACRIOCISTITE NEONATAL	7716	F	60000000000000000552	27000000000001600
8	70000000097640000552	INFECCAO NEONATAL POR CANDIDA	7717	F	60000000000000000552	27000000000001600
9	70000000097650000552	INFECCOES ESPECIFICAS NCOP DO PERIODO PERINATAL	7718	F	60000000000000000552	27000000000001600