LAPPEENRANTA-LAHTI UNIVERSITY OF TECHNOLOGY LUT
School of Engineering Science
Industrial Engineering and Management
Business Analytics
Master’s Thesis

Elli Saarenmaa

ASSESSING AND MEASURING DATA QUALITY IN CREDIT RISK MODELLING

Supervisors: Professor Pasi Luukka
Research Fellow Jan Stoklasa

ABSTRACT

Author: Elli Saarenmaa

Title: Assessing and measuring data quality in credit risk modelling

Year: 2019 **Place:** Helsinki

Master's Thesis. LUT University, Industrial Engineering and Management.
92 pages, 2 figures and 14 tables
Supervisors: Professor Pasi Luukka and Research Fellow Jan Stoklasa

Keywords: data quality, information quality, data quality assessment, data quality metrics, credit risk modelling

Hakusanat: datan laatu, tiedon laatu, datan laadun arviointi, datan laadun mittaaminen, luottoriskimallinnus

Due to increasing regulatory demands, banking institutions are adopting better data quality management practices. The aim of this thesis was to present an overview of the methods to assess and measure data quality in credit risk modelling. The thesis presents the regulatory requirements that need to be complied with. Then, data quality assessment and measuring methods were discovered by conducting a literature review. The final analysis consisted of 44 items. The results of the literature review are analyzed and applied to a case data obtained from a case company. The findings of this thesis could be used to improve the data quality management practices and processes.

The regulation requires banks' to have consistent criteria and metrics with clearly set tolerance levels. Banks should apply the presented assessment practices based on their own internal data and models. The results of this thesis show that the implementation of data quality assessment methods require collaboration of experts from different fields. It is necessary to understand the use case of data and what the data represents when assessing data quality. The acceptable quality thresholds should be defined so that the undetected errors do not have significant effect on the ratings and the metrics should be chosen so that the results are efficiently obtained. Further study is required to obtain the quality thresholds and best applications of methods in the credit risk modelling context.

TIIVISTELMÄ

Tekijä: Elli Saarenmaa

Työn nimi: Datan laadun arviointi ja mittaus luottoriskimallinnuksessa

Vuosi: 2019 **Paikka:** Helsinki

Diplomityö. Lappeenrannan-Lahden teknillinen yliopisto LUT, Tuotantotalous.
92 sivua, 2 kuvaa ja 14 taulukkoa
Tarkastajat: professori Pasi Luukka ja tutkijatohtori Jan Stoklasa

Hakusanat: datan laatu, tiedon laatu, datan laadun arviointi, datan laadun mittaus, luottoriskimallinnus
Keywords: data quality, information quality, data quality assessment, data quality metrics, credit risk modelling

Lisääntyneen säätelyn vuoksi pankit kehittävät datan laadunhallintakäytäntöjään. Tämän diplomityön tavoitteena oli antaa yleiskuva menetelmistä, joilla datan laatua voidaan arvioida ja mitata luottoriskimallinnuksessa. Ensimmäiseksi diplomityössä esitetään sääntelyvaatimukset, joita on datan laadun osalta noudatettava. Seuraavaksi datan laadun arviointi- ja mittausmenetelmiä kartoitettiin toteuttamalla kirjallisuuskatsaus. Lopulliseen analyysiin sisältyi 44 julkaisua. Lopuksi kirjallisuuskatsauksen tulokset analysoidaan ja niitä sovelletaan case-yritykseltä saatuun dataan. Diplomityön tuloksia voidaan käyttää tiedon laadunhallintakäytäntöjen ja -prosessien kehittämiseen.

Sääntely edellyttää pankeilta johdonmukaisia kriteereitä ja mittareita ennalta määrätyillä kynnysarvoilla. Pankkien tulisi soveltaa työssä esitettyjä arviointimenetelmiä sisäiseen dataansa ja malleihinsa. Diplomityön tulokset osoittavat, että datan laadun arviointimenetelmien toimeenpano edellyttää eri alojen asiantuntijoiden yhteistyötä. Datan laatua arvioitaessa on ymmärrettävä datan käyttötarkoitus ja tiedon sisältö. Hyväksyttävä laatutaso tulisi määritellä siten, että havaitsemattomilla virheillä ei ole merkittävää vaikutusta luottoriskimallinnuksen tuloksiin. Sovellettavat arviointimenetelmät tulisi valita siten, että tulokset saadaan tehokkaasti. Lisätutkimusta tarvitaan laatutasojen ja parhaimpien käytäntöjen määrittämiseksi luottoriskimallinnuksessa.

CONTENTS

1 INTRODUCTION

Data has become an important asset in almost any industry. Financial services companies are no longer traditional money businesses since their business models depend largely on information. Information requires data, and in order to respond to regulators' and customers' higher demands, banking institution use diverse data. Data is seen as strategic asset that can be used to gain competitive advantage, and to achieve greater growth and profitability. (Robert Morris Associates 2017)

While information has become more essential for business and the amount of data collected and managed has grown, data quality has become an important topic. Since information systems have advanced rapidly, appropriate quality controlling tools have not necessarily been adopted. (Olson 2003) During the past years, most banking organizations have been taking initiatives for improving data quality (Robert Morris Associates 2017). Yet there still remain opportunities for further enhancements. Olson (2003) argued data quality issues are usually managed reactively rather than proactively. Continuous proactive data quality management would allow companies to gain additional value since higher quality data helps them to make better decisions. (Olson 2003)

This research is focusing on credit risk modelling data in financial industry. Even though the topic of data quality has been discussed during the last years there are no consensus on the measures that should be used to assess data quality. Banking industry faces additional challenges due to increasing risk data aggregation regulation. The study covers regulatory requirements for credit risk modelling data and examines the methodologies to assess and measure data quality based on those requirements. The findings of this study can help financial institutions to improve their data quality management and assessment frameworks.

1.1 Background

Data quality has gained increasing attention among scholars and regulators over the last years (Baesens et al. 2013). This is due to the growing amount of data companies need to collect and manage. Data quality issues are universal, but the best-practices of data quality measures and improvement actions cannot be defined universally. Data quality issues occur universally among large organizations whereas many organizations presume that their data quality is adequate. Data quality can become an issue even when the same database has been used for years without problems since new use cases that have higher requirements are presented. The level of acceptable data quality cannot be universally defined since it depends on the intended use of data. (Olson 2003) The banking industry is faced with increasing demands due to the speed and complexity of the world's financial markets, tightening regulation, rapid advances and declining costs in technology, and customers' growing demand for fast and high-quality service (Robert Morris Associates 2016). Financial institutions collect diverse data which causes challenges for getting accurate and timely data in the right format to use organization-wide.

The global financial crisis that began in 2007 showed there is a constant need for ensuring risk data aggregation and risk reporting are integrated into all risk activities (Bank for International Settlements 2013). The Basel Committee on Banking Supervision sets global standards for the prudential regulation (Bank for International Settlements 2019), the European Union introduces legal acts (European Union 2019), and the European Banking Authority prepares regulatory and non-regulatory documents (European Banking Authority 2019b) for banking institutions. The legislation and guidelines of these institutions form the basis of data quality requirements. The internal models and the data used in the models are supervised by European Central Bank that authorizes the use of internal models for credit risk modelling. (European Central Bank 2018a)

Banks have recognized the need for continued improvement of data. The RMA/AFS survey examines how data quality is perceived and managed by banks internationally. For regulatory reasons, ensuring data quality has been a primary concern to almost every bank. Majority of respondents in 2016 all over the world noted that their organizations have developed or changed their data management strategies over the last three years. Many organizations have increased the number of staff and taken short-term clean up initiatives to improve data quality. The survey suggests that banks have noticed improvement towards better data quality management on the enterprise level but consensus on how to govern and manage data quality policies has not yet been achieved. The respondents pointed out organizational silos, technology limitations, other objectives, and budgets as the main obstacles. (Robert Morris Associates 2017) Even though banks have been working towards improving the quality of their data, there remains a lot to be done in order to gain competitive advantage. In 2016, European Central Bank (2018) launched a thematic review to assess how well 25 institutions complied with Basel risk aggregation requirements. Their review showed the implementation status was unsatisfactory. Mainly the deficiencies were due to the institutions lacking clarity on the roles and responsibilities of data quality management. The supervisors believed full implementation of the practices will be an ongoing process that will still last for a few years. (European Central Bank 2018b) Thus yet today, the topic of data quality management is very relevant in the credit risk modelling.

1.2 Objectives and scope

This thesis aims to help financial companies to find ways for improving their data quality assessment practices. There exists a vast amount of literature discussing data quality and the techniques and methodologies to address data quality issues. Yet, new credit risk modelling regulation causes uncertainties on the best practices since no general practices have not yet been adopted. There is a need for better understanding of the linkage between the data quality assessment techniques and the regulatory requirements. The objective of this thesis is to identify the data quality assessment methodologies that could be used in the credit risk modelling to comply with the risk data regulation. The intended outcome of this study is to get an overview on the methods that could be used to assess and measure data quality, and to propose methods for assessing, measuring and quantifying data quality in credit risk modelling. The research question was formed as:

Which methods could be used to assess and measure data quality in credit risk modelling to comply with the regulatory needs?

This research is limited to short-term data quality assessment in data warehousing environment since it is directed to a line of business rather than organization-wide. Long-term improvements would require covering system design development, processes for tracking the source of the quality issues and procedures for organization level data quality program which are out of the scope of this study. The study will only include input data of risk modelling process and not address the quality of risk modelling process output data (the final ratings). The study focuses on the requirements for internal ratings-based (IRB) models.

1.3 Methodology

The theoretical framework of this thesis presents the requirements of data quality in the credit risk modelling context. Next, the study was carried out by conducting a literature review on data quality assessment methods given the regulatory requirements. As there are yet no generally accepted practices for data quality assessment in the banking field, literature review was chosen as the research method to provide an overview on the existing practices in different fields. It was not necessary to collect and identify all evidence but rather discover different ways to address the issue of data quality assessment. Since all the methods presented are not from the banking field, the results only give indication on the possible assessment methods that could then be adjusted for specific use. The findings of the literature review could be used to improve the data quality management processes and metrics.

Relevant literature was searched from university's library sources, Web of Science and Scopus. The following key words were searched from title and abstract in different combinations: "data quality" or "information quality", assess* or measur* or valid* or examin*, completeness or accuracy or consistency or timeliness or uniqueness or validity. After the suitable articles for the purpose were chosen, their references were scanned through to find more studies on the subject. The final literature review analysis consisted of 44 papers. The data quality assessment techniques and metrics used are summarized from the researched articles. The objective is to identify assessment and measuring techniques that could be used for risk modelling data quality in data warehousing environment.

Finally, the results from the analysis of assessment and measuring technique are applied to an empirical case study. The data consisted of credit agreement data collected at three different points of time from different years. The data was constructed to include date type, string type, decimal type, binary type and character type attributes. There were 11 attributes and around 1.85 million records in total. A data set was collected specifically for the purpose of the analysis and does not represent the real data set used for modelling purposes. Industry professionals' input was used to distinguish the methods that could be applied to the data.

1.4 Structure

This thesis contains six chapters and proceeds as follows. The first chapter presents the background and the objectives for the study. Data quality is described in chapter 2. It presents the definition of data quality and what it means for data to be perceived as high quality. It discusses the process of how data quality should be assessed. Chapter 3 then discusses the regulatory requirements of data quality assessment for risk modelling purposes. In addition, the chapter discusses the background of internal ratings based approaches regulation and presents the process of credit risk modelling. Finally, the chapter discusses what requirements credit risk modelling regulation impose on data quality. Chapter 4 presents the research problem and present the results of the literature review conducted. The assessment and measuring techniques are presented in detail. Lastly, the results are summarized and analyzed. Chapter 5 applies the techniques to a case study data. To conclude, chapter 6 discusses and summarizes the results of the thesis.

2 UNDERSTANDING DATA QUALITY

Data quality has been discussed for many years, yet the topic has become increasingly important during the last years. Batini and Scannapieca (2006) conceive data quality as a multidisciplinary concept as the topic has been discussed by researchers of multiple fields. Data quality has been examined by fields such as statistics, management, and computer science since the 1960's. Computer scientists started to discuss database quality assessment and improvements in the 1990's. (Batini and Scannapieca 2006)

The definition and importance of data quality seems to be widely agreed upon. This chapter focuses on defining what is meant by data quality and what it means for data to be of high quality. Olson (2003) argues that it is never possible to reach perfect quality, yet high-quality data can still be distinguished from poor quality data. The chapter introduces how the level of quality is defined. Next, the chapter introduces the dimensions of data quality to better understand the elements that data quality consists of. Thirdly, the chapter focuses on why data quality issues occur in the first place and why the issue of quality is important. Lastly, the chapter presents how the major quality problems can be addressed and what phases should the data quality assessment process include.

2.1 Defining data quality

Many scholars agree on the definition of data quality. Wang and Strong (1996) affirm the term data quality is widely defined from the viewpoint of the consumer and defines data quality as "fit for use" by data consumers. Ballou and Tayi (1998) agree with this definition and state that the definition of high-quality data is relative. Olson (2013) similarly defines data having quality if it fulfils the requirements for what it is intended to be used for, and lacking quality if it doesn't fulfil the requirements. The intended use of data can thus be seen as a component of data quality. Baesens et al. (2013) state the objective of credit risk modelling is to identify as accurately as possible the credit risks resulting from possible defaults on loans. Thus in the case of credit risk modelling, high quality data is thus defined as data that allows the modelers to reach accurate modelling outcomes.

Data quality can be defined using a second component. Sebastian-Coleman (2013) defines data quality by two factors, the first being the same as previously mentioned. According to Sebastian-Coleman, data quality is defined by the level it meets the expectations of the users (how well it suits the intended use) and the level it represents the objects or events it is supposed to represent. (Sebastian-Coleman 2013) In this thesis, the two-component definition is used. In the credit risk modelling context, data quality is thus defined as suitable for the credit risk modelling and representing the objects or events it is supposed to be representing.

Data quality is many times considered as part of information quality. Information is defined as data with meaning or purpose. (Sebastian-Coleman 2013 p. 14) Baesens et al. (2010) also state information quality is sometimes used in the literature to refer to data quality. Data and information then form the basis of knowledge (Sebastian-Coleman 2013 p. 14).

2.2 Defining data quality dimension

It has been agreed for long that data quality is best represented by several dimensions (Ballou and Tayi 1998). Wang and Strong (1996) define data quality dimensions as data quality attributes that can be regarded as a single construct of data quality. Data quality is often misinterpreted to the concept of data accuracy. Data quality consists of several dimensions and accuracy is a sub concept of data quality. (Batini and Scannapieca 2006) Related research on the topic present several dimensions which are named in this section.

Dimensions can be clustered, and some dimension can be presented as subdimensions of other dimensions. Wang and Strong (1996) list the most important dimensions based on their literature review as accuracy, timeliness, precision, reliability, currency, completeness, relevancy, accessibility and interpretability. They studied how data consumers see data quality, and 179 attributes were identified in their survey. They then surveyed the importance of 118 of those dimensions and most of them were identified as important. Finally, the dimensions were categorized under four clusters of intrinsic, contextual, representational and accessibility data quality. The four clusters included 15 dimensions in total. (Wang and Strong 1996)

Batini and Scannapieca (2016a) divided the dimensions into eight categories. The first category was represented by accuracy, and included correctness, validity, and precision. The second category of completeness included pertinence and relevance. The third represented by redundancy had minimality, compactness, and conciseness included. The fourth included comprehensibility, clarity, and simplicity, and was represented by readability. The fifth category was accessibility with availability included. The sixth category was consistency which consisted cohesion, and coherence. Usefulness was the seventh category. The eighth and last category named was trust which included of believability, reliability, and reputation. They named accuracy, completeness, currency, and consistency being the most important categories. (Batini and Scannapieca 2016a p. 23)

Many researchers agree on the most important dimensions. Ballou and Pazer (1985) described four dimensions: accuracy, completeness, consistency, and timeliness. Olson (2003) states that data must be accurate, timely, relevant, complete, understood, and trusted in order to comply with the intended use. Sebastian-Coleman (2013) discusses six dimensions of data quality which include completeness, timeliness, accuracy and validity, consistency, and integrity. Cappiello et al. (2018) discuss three dimensions in their study: accuracy, completeness and coherence/consistency. The dimensions of accuracy, timeliness, completeness and consistency are agreed by most of the researchers. The suitable dimensions for risk modelling data quality assessment purposes are later introduced and discussed in more detail in the later chapters.

2.3 Reasons behind data quality issues

Despite the fact that data is an important asset for many companies, their databases contain large amounts of poor-quality data. Data quality issues occur universally among large organizations even though companies tend to withhold the knowledge of data quality problems. It is not a matter of inadequate management style but rather a result of the rapid development of information systems lacking the advancements in quality controlling tools. Companies have implemented new IT systems within a short period of time and have not had the capabilities to monitor data quality. Since new practical applications for data are presented, data quality can become an issue even when the same database has been used for years without problems. This happens when new use cases have higher demands for data. (Olson 2003) Ballou and Tayi (1998) underline that data quality is a wide problem since the values can be wrong in several different ways. Data could be high-quality for most of the dimensions but deficient on a critical few. (Ballou and Tayi 1998) Olson (2003) adds some incorrect values are not likely to cause harm but cumulative result of multiple incorrect values can change the outcomes drastically.

There are many reasons behind the poor data quality of companies. The authors of the RMA/AFS survey emphasize that the greatest data quality problems have remained the same for a decade of the survey’s history. The top issues in the 2007 survey (in descending order) were quality, information silos, data entered multiple times, IT challenges and costs. Four of these issues have remained the same as the top problems in 2016. Those were (in descending order) information silos between lines-of-businesses, IT challenges, data entered several times in several systems, lack of consistent data definitions and costs. The authors argue that banks have not yet taken appropriate actions in solving the constantly recurring data quality issues. (Robert Morris Associates 2017) Olson (2003) adds that the requirements are usually poorly articulated, the data acceptance testing of systems is poorly designed, and the data creation processes are inadequate.

The data transaction paths are typically complex, and it is not unusual for a company to have several application server providers. Since the system architecture is complex, there are multiple ways for incorrect values occurring. Olson lists four main areas where poor-quality data is likely to occur: initial data entry, data decay, moving and restructuring, and using. Initial data entry means invalid data is entered in the first place by mistake, the process is confusing or poorly designed, wrong values are entered on purpose, or system error has occurred. Data decay means data is originally correct but becomes incorrect through time. When data is used, it should be understood and easily accessible in order for it to be used for a specific purpose. (Olson 2003)

Systems are constantly changing due to changing needs and the information in a database is often gathered from several source systems. The changes can lead to inconsistencies: the way information is recorded is altered, or subdivisions to possible values are added. Trend analysis based on information that has changed over time will lead to wrong conclusions. Also, different user groups may insert or delete data with different criterions. In order to achieve the quality level needed, the changes should be made all the way to the source. In addition, some data quality issues are more likely to be located since the use case affects the chance of recognizing incorrect values. Some data is always more important than other data, thus some data issues tend to be corrected immediately when needed. If the changes are not documented, the users are unaware of the current state of data quality and the data might not be corrected in all of the systems. (Olson 2013)

Ballou and Tayi (1998) state that low priority given to data management causes poor-quality data. Even though managing data and monitoring data quality is considered as an important activity, it is not top priority for management. Thus, not enough budget is allocated for improving data quality. (Ballou and Tayi 1998) This might be changing as data is considered increasingly important to companies' business strategies. Especially in the financial services industry, regulation increases pressure on funding data quality governance.

2.4 The importance of data quality

High-quality data is a necessity for information quality. Information quality helps companies to gain competitive advantage. Olson (2003) argued executives may not be aware of the potential value of fixing data quality issues. Funk et al. (2006) agrees that executives are usually unaware of the issues or they might believe the IT department can address it. Lee argues it is critical for employees of different levels to understand the importance of data quality before data quality can be achieved. Understanding the importance promotes active participation in data quality management processes. (Funk et al. 2006)

Funk et al. (2006) lists four major reasons for the importance of data quality: high-quality data is a valuable asset, it increases customer satisfaction, it improves revenues and profit, and it can bring strategic competitive advantage. In contrast, poor data quality has serious impacts on firm's effectiveness. Data quality issues are costing firms a great deal of money. Data quality experts have estimated poor-quality data to cost organizations as much as 15–

25 % of operating profit. (Olson 2003) Eppler and Helfert (2004) listed and categorized costs of poor quality data. They concluded low quality data directly causes costs such as costs of verification, processing, distributing, tracking, training and repairing. Indirectly it causes costs for example due to data loss, customer loss, lower reputation, wrong decisions taken and missed opportunities. (Eppler and Helfert 2004) Batini and Scannapieca (2006) argued that poor data quality affects organizations negatively every day but the issues are not necessarily traced to data quality. They agreed data quality has significant consequences on the efficiency and productivity of businesses. (Batini and Scannapieca 2006) Also Olson (2003) emphasizes poor data quality causes organizations financial losses, waste of time, incorrect decisions and missed opportunities. He agrees many organizations believe their data quality is adequate, thus they are unaware of the extent of the losses and miss the opportunity to improve their efficiency. Marsh (2005) collected effects of poor data quality from the reports done by Gartner Group, PWC, and The Data Warehousing Institute. According to the reports in 2005, every year in the US over 600 billion dollars were lost only due to poorly targeted mailings and staff expenses. They additionally stated that due to poor quality data, 88 percent of data integration projects were exceeding budgets extensively or had totally failed, and 33 percent of companies had postponed or abandoned new IT systems. (Marsh 2005) Baesens et al. (2013) states poor-quality data have an impact on customer satisfaction, it causes extra operational costs, and can lower employee job satisfaction. Most importantly, it may cause inaccurate credit decisions, which is an important aspect in credit risk management (Baesens et al. 2013).

Improving data quality for modelling purposes improves the accuracy of the final results, thus improving the credit approval decision-making (Baesens et. al 2013). Financial scoring process takes internal and external data as input and follow a series of activities with pre-defined rules, resulting a credit rating for the customer. Poor-quality input data or poor-quality preprocessing can result in poor-quality output data, in other words poor-quality ratings. The quality of input data should be evaluated in the beginning of the process and the quality of the data manipulation should be assessed to meet the sufficient quality level. (Cappiello et al. 2018) It could thus be stated that high-quality data is important in order to accomplish high-quality credit risk modelling results. Baesens et. al (2013) emphasize imprecise calculations and estimates of credit risk parameters can result in financial losses, or to a greater extent, even bankruptcy of the institution. Thus, the importance of data quality in credit risk modelling is undisputed.

2.5 Achieving high-quality data

As discussed earlier, data quality issues are usually managed reactively rather than proactively. In order to achieve improvements, executives should take proactive steps towards data quality management. It requires that managing data quality is done continuously in the long-term. Olson remarks that organizations should invest in system design and continuous monitoring of data collection and take aggressive actions to solve issues that generate inaccurate data. Short-term improvement of data quality can be achieved by filtering of input data, cleansing of database data, and creating awareness of quality for the end-users. (Olson 2003)

Data quality improvements need the cooperation of the whole organization to establish coherent policies organization-wide. Solving the quality issues at the source requires collaboration of the executives from the top of the organization and the targeted technology and process initiatives from below (Robert Morris Associates 2017). Funk et al. (2006) agrees the highest-level executives should be part of the change and awareness of the issues needs to be reached before an organization can truly improve its data quality. Intuitive description of the state of quality is not enough but organizations need realistic and usable policies. Companies should measure both subjective and objective variables of data quality. (Funk et al. 2006) The rules for inserting and deleting data should be clearly defined organization-wide. Additionally, organizations should have clearly defined and documented during what periods has data been recorded consistently. (Olson 2003)

As the definition of data quality showed, high-quality data is a relative matter. Data quality assessment cannot be done without specific information on the intended use cases. When databases are built, the requirements for the use should be gathered first and the design should be assigned to those requirements. Yet in business context, not all use cases are known or defined at the time database was designed. In order to address the issue, the implementations need flexibility. (Olson 2003)

Olson (2003) states perfect data accuracy can never be reached but it is still possible to distinguish between high and poor quality, and it is possible to get correct data to a degree that it is highly useful for the use cases. For example, a database with 0,5% inaccuracy level would probably be considered as high quality by most users. Tolerance level should be chosen so that the application of data provides high quality decisions which would not largely change even if the data was 100% correct. (Olson 2003) In credit risk modelling context, Cappiello et al. (2018) suggest that new data quality controls should be implemented, or the existing ones should be improved if poor data quality has high effect on the ratings. If poor data quality has low effect, they suggest reacting to issues when they occur rather than implementing new data quality controls.

2.6 Data quality management process

Many scholars discuss the process steps for data quality management. Baesens et al. (2013) state most programs include four process which are data quality definition, measurement, analysis and improvement. Also, Cappiello et al. (2018) identified four stages for data quality management which include defining the data quality dimensions, measuring the chosen dimensions, analyzing the root causes of data quality issues, and finding improvements. In this research, this approach is applied. First, the dimensions to be assessed are identified and defined. Then, the most important measures are found through literature review, and lastly the data is analyzed. The scope of this research does not include data quality improvement.

Baesens et al. (2013) and Cappiello et al. (2018) named data quality definition as the first step where the appropriate dimensions should be identified. Granese et al. (2015) suggest that the quality assessment starts with choosing the appropriate data quality dimensions. The chosen (most important) attributes are measured and given scores against each of these dimensions. The data quality assessment is conducted using business rules and data profiling. The quality scores could be aggregated at function or enterprise level. (Granese et al. 2015) Batini and Scannapieca (2006) agree that data quality assessment process should start from selecting the dimensions to be measured. Cappiello et al. (2018) in contrast argue that identifying the data quality dimensions and their control methods is mostly done by experts who might be biased. The control methods are therefore adopted on the basis of expected and evident data quality problems and they lack the effectiveness in dealing with unobserved problems. Thus, the result of quality values and the monitored dimensions could be overestimated or underestimated on the process outcome. (Cappiello et al. 2018) In this thesis, the most important dimensions are chosen based on the regulatory demands. It is not discussed whether other dimensions should be included as well.

Granese et al. (2015) also suggest that the most important attributes for the specific business area should be identified in the beginning of data quality assessment process. In their opinion, the size and complexity of data population of a large financial institution makes complete data quality assessments for all the attributes impractical. Thus, the required attributes being measured should be identified as well. (Granese et al. 2015)

Olson (2003) argued that most incorrect values can be identified if enough effort is devoted for searching them. There are two types of options for finding incorrect data: reverification and analysis. Reverification means manually starting to track information from the original source and check every value. Not all the errors could be identified since wrong values could be inserted again in the reverification process. In real life, this process is excessively time consuming and expensive for most organizations. Additionally, reverification process is not always possible for all data if the data does not exist in the source systems anymore. As a monitoring process before data use, it would most definitely violate the timeliness requirements. Selective reverification could be used as a monitoring technique so that only a small sample of records are reverified. (Olson 2003) To conclude, even if most of the incorrect data could be identified, it is not always economically feasible, and it is a trade-off between timeliness requirements. Heinrich et al. (2018) agree that inadequate measuring could lead to excessive costs. The metrics applied should be economically efficient to use them in practice (Heinrich et al. 2018). Even if errors could be best discovered through manual inspection of values, it is so time-consuming that it does not make sense when datasets are immensely large. Thus, the best data quality metrics identifies as many quality issues as possible in a least amount of time.

3 CREDIT RISK MODELLING

The aim of credit risk modelling is to determine the regulatory capital needed to compensate for potential losses (Baesens et al. 2010). The regulation requires banking institutions to evaluate the credit risks they've invested on. The aim of the institutions is to identify as accurately as possible the credit risks resulting from possible defaults on loans. (Baesens et al. 2013) The global financial crisis that began in 2007 showed some banks were not able to adequately aggregate their risk exposure. Regulators then increased the requirements to ensure risk data aggregation and risk reporting are integrated into all risk activities. Risk data aggregation means the activities to define, collect and process risk data to comply with the bank's risk reporting requirements and become able to quantify their risk tolerance. (Bank for International Settlements 2013) The regulation affects the capital requirements and solvency of financial institutions. Regulators have increased their attention also to data quality issues on credit risk management since the modelling is based on banks' internal data. Data quality is thus closely monitored in credit risk modelling. (Baesens et al. 2013).

Managing data quality is essential for meeting the regulatory demands. Prorokowski and Prorokowski (2015) remark the financial industry is rapidly becoming more regulated thus financial institutions should concentrate on developing their risk data aggregation processes. They argue banks need to implement new tools and find efficient ways to achieve high standards. New regulation requires banks to improve their data aggregation processes, and to establish clear frameworks. The improvements would allow banks to remedy more easily from future episodes of financial distress. (Prorokowski and Prorokowski 2015) Gupta and Kulkarni (2016) show data quality issues can have notable impact on key risk numbers and cause inaccuracies in risk reports. Inconsistencies in data structures and formats, and the absence of common data systems and terminology across companies cause challenges for risk data aggregations. They name identifying data quality problems and understanding the root causes of them as critical part of complying with the regulatory requirements. (Gupta and Kulkarni 2016)

For regulatory reasons, insuring data quality has been a primary concern to almost every bank. Many organizations have increased the number of staff and taken short-term clean up initiatives to improve data quality. Banking institutions are also developing their data quality frameworks in order to respond to the regulatory demands. (Robert Morris Associates 2017) The aim of the case company is to establish an effective data quality management framework to be part of their credit risk modelling projects.

This chapter first presents what are internal ratings based models and how the modelling projects are carried out. This chapter then focuses on explaining the history of regulation for internal ratings based approaches and the requirements for data quality posed by European Central Bank (ECB). The focus is on the components and dimensions of data quality that need to be assessed and monitored but the chapter also touches upon the subject of requirements of data quality management framework, responsibilities and reporting in order to get a comprehensive understanding of data quality management. The IT system requirements are not included in this study.

3.1 Credit risk modelling process

The capital requirements make sure financial institutions are able to compensate for possible losses at the 99.9 % confidence level (Baesens et al. 2010). The modelling process is primarily concerned with quantifying the losses caused by obligors' failure to repay loans (Baesens et al. 2013). The capital needed is determined as a percentage of the risk weighted assets (RWA) (Baesens et al. 2010). The formula of risk weighted assets is presented in the Capital Requirement Regulation (EU/575/2013, article 153) and it is given as

$$RWA = RW \cdot EAD \tag{1}$$

where EAD represents the exposure at default and it is estimated using the conversion factor (CF) parameters, and the risk weight RW is a function of the parameters probability of default (PD) and loss given default (LGD). Detailed information on the calculation of risk weighted assets can be found from the Capital Requirement Regulation. Financial institutions can use their own best estimates of the PD, LGD and CF or the values given by the regulator (EU/575/2013). Bank's regulatory capital can thus be established by using a standardized approach where the parameters are given or by using the internal ratings based approach where the parameters can be estimated by the bank. The parameters are estimated using bank’s own internal data. IRB approach increases administration costs, but it allows the institutions to have lower capital requirements thus they are often used. (Rutkowski and Tarca 2016) This thesis focuses on the process of IRB approach modelling.

The credit risk models can include data on account information of the loan and the loan applicant (Baesens et al. 2013). For data quality assessment, it is important to understand the use case of data. Thus, the process of credit risk modelling in the case company is presented in the figure 1. The process is described on a high-level and does not include all the essential tasks performed during the modelling process. The results are analyzed at each phase and at any phase, there can be a return to the previous phases if changes are needed.

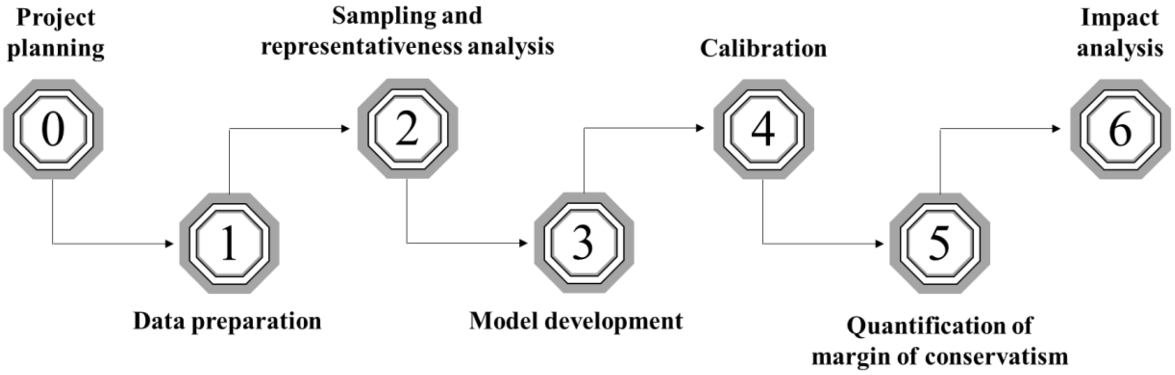

Figure 1. The process of credit risk modelling in the case company

Before a modelling project can begin, the objective of the project is identified and clarified. The necessary project tasks and their timeline are identified and planned. Once an overview has been obtained, a final decision is made whether to proceed with the proposed plan. After the initiation phase, data is prepared and analyzed for the project. Data preparation phase ensures data is comprehensive and usable in the specific use case. The second phase consists of preparing data samples needed for different stages of modelling and ensuring that the samples are correctly representing the portfolio.

Data quality assessment is conducted during the data preparation phase. Additional analysis such as representativeness analysis on the quality of data is conducted during other phases of the modelling project as well. For this thesis only the analysis of completeness, accuracy, consistency, timeliness, uniqueness and validity is considered.

Once all the preparation is done, the model is developed. The model is chosen so that it has a good predictive power and is reasonable for the business. Next, the model is calibrated to ensure appropriate levels of risk parameters. Then, the strengths and weaknesses of the model are analyzed, and the deficiencies are identified. Since a model can never be a perfect representation of the future events, estimation of the uncertainty is needed. Basel II framework requires banks to estimate and add a margin of conservatism (MoC) to reflect model errors and uncertainty (De Jongh et al. 2017). MoC is quantified based on the deficiencies of the model and the data used. Finally, the impact of the model is analyzed. When the modelling project is finalized, approval process is run if supervisory approval is needed. When the model is approved, the model is implemented and brought to production.

3.2 Regulation of internal models

The Basel Committee on Banking Supervision (BCBS) primarily sets global standards for the prudential regulation of banks. The BCBS consists of 45 members which include central banks and bank supervisors from 28 jurisdictions. (Bank for International Settlements 2019) European Union (EU) introduces different types of legal acts. Regulations set by EU are binding legislative acts that must be applied everywhere in EU. Directives set the objectives EU countries need to fulfill but the countries can set their own national laws to address how these objectives are reached. (European Union 2019) European Banking Authority (EBA) then prepares regulatory and non-regulatory documents such as Technical Standards, Guidelines, Recommendations, Opinions and ad-hoc or regular reports. The Binding Technical Standards are based on EU Directives or Regulation, and they are legal acts that make specifications to EU legislation. The objective of the EBA is to ensure prudential requirements are consistently applied by providing supervisory practices. In addition, EBA is mandated to analyze risks and vulnerabilities in the EU banking sector. (European Banking Authority 2019b) Finally, European Central Bank (ECB) is the authority who supervises and authorizes banks for the use of internal models for credit risk (European Central Bank 2018a). ECB supervises directly significant entities which in Finland consist of Nordea Bank Abp, Kuntarahoitus Oyj, and OP Osuuskunta based on their total assets. The other entities are supervised by Finanssivalvonta, and indirectly by ECB. (European Central Bank 2019)

The Basel Committee on Banking Supervision first introduced the Internal Ratings Based approach in Basel II framework in 2006. The IRB approach allowed banking institutions to impose their own capital requirements based on risk parameter estimation. Banking institutions could estimate the risk parameters for their own organization. In Europe, Internal Ratings Based Approach was introduced by the Capital Requirements Directive in 2006. (European Banking Authority 2019a)

During the latest financial crisis, some banks lacked the abilities to manage their risk exposure which had severe consequences to them but also to the whole financial system. Regulators found that banks' internal models and their supervision was inadequate. In 2010, Basel Committee published the Basel III framework to strengthen capital and liquidity standards. (Bank for International Settlements 2011) Regulators acknowledged there is a need for improving banks' risk data aggregation processes. In January 2013, the Basel Committee on Banking Supervision published the principles for effective risk data aggregation and risk reporting (BCBS 239). The principles require banks to develop risk data aggregation framework to prepare for possible issues beforehand. The banks should develop their risk data management abilities over the long term. The Basel Committee stated the future benefits due to faster and better information sharing and decision making would compensate for the required investment costs. The principles also concern internal ratings-based approaches for credit risk modelling. (Bank for International Settlements 2013)

In Europe, the Capital Requirements Directive of 2006 directive was replaced in June 2013 by Regulation (EU) No 575/2013 known as Capital Requirements Regulation (CRR) and Directive 2013/36/EU known as Capital Requirements Directive (CRD). The CRR assigned the European Banking Authority (EBA) to provide clear technical standards and guidelines to secure the IRB requirements are consistently applied. The EBA have guidelines with the purpose to clarify risk parameter and own funds requirements. Their purpose is to reduce risk parameter variability to achieve comparability. They have published guidelines on PD and LGD estimation. (European Banking Authority 2017) The EBA have additionally published final draft regulatory technical standards (RTS) on the IRB assessment methodology for the validation of the models (European Banking Authority 2019a).

The revised and finalized Basel III framework was published in December 2017 while it was expected to be published earlier. The implementation of the framework should be done before 1 January 2022. (European Banking Authority 2019a) Due to new requirements, new models are adapted in the case company and data quality is increasingly considered as part of the modelling projects. As the regulation and supervision is relatively new, there exists no generally accepted practices on the assessment and measuring techniques in the field.

European Central Bank (ECB) is the authority who authorizes banks for the use of internal models for credit risk. ECB have published a guide on how the requirements are understood and applied based on the current applicable EU and national laws. The legal background of their data quality requirements is based on the CRR and the EBA guidelines on PD and LGD. Their data quality requirements additionally reference on the final draft RTS on assessment methodology for IRB, and BCBS 239. (European Central Bank 2018a) The data quality requirements in this thesis are based on these publications.

3.3 Data quality requirements

Prorokowski and Prorokowski (2015) argued data collection, integration processes and validation that support risk management and regulation keep posing technical challenges to banks. Many times, risk aggregation processes demand great manual effort. In addition, banks are lacking transparency over risk data governance. The BCBS 239 standards require banks to systematically revise their current data issues. Nevertheless, they emphasize that to comply with the principles, the actions of fixing the current data errors and filling the missing values are not enough. The BCBS requirements recommend banks to establish effective risk data governance and IT systems. Ideally, the new established standards would improve the understanding of risk data across the whole organization. Essentially, risk management processes require making the best decisions with the available information. (Prorokowski and Prorokowski 2015)

Prorokowski and Prorokowski (2015) state according to the BCBS 239 principles each data set is supposed to be easily traced to its source and validated for being able to compare the values across different source, vendors or legal entities. Bank for International Settlements (2013) lists 14 principles in total for which four principles are for risk data aggregation capabilities: accuracy and integrity of risk data, completeness, timeliness and adaptability. They state banks should be capable of providing accurate, reliable and complete risk data which is largely automatically aggregated. The data should be able to have risk data captured and aggregated across the banking group. Aggregate risk data needs to be available to all relevant stakeholders in a timely manner. Banks should be able to provide aggregated risk data requested by supervisors on-demand. (Bank for International Settlements 2013)

European Central Bank (2018a) states that banking institutions should employ solid data quality management practices in order to provide sufficient support for its credit risk management purposes. They underline institutions should deploy data quality practices and processes at group level. Companies should set and administer an effective framework which is applicable to both internal and external data in the modelling related processes. For the framework to be comprehensive, it should include governance principles, description of the scope, consistent criteria and metrics, continuous assessment procedures, sufficient reporting, and it should cover all relevant data quality dimensions. (European Central Bank 2018a)

In order to comply with the governance principle requirements, the framework should be current and revised periodically, it should be approved by senior executives, and verified regularly by independent auditing unit. Responsibility for the governance should be clearly divided throughout the institution to the appropriate staff members. The scope of the framework means it should include all relevant data quality dimensions, and the complete lifecycle from data entry to reporting. The framework should consider both historical data and recent up-to-date databases. ECB underline that data quality standards should be set to all stated dimensions for all modelling input data and for each stage of the data life cycle. (European Central Bank 2018a)

European Central Bank (2018a) understands data quality dimensions are important part of data quality management framework for complying with the regulatory requirements. They require effective data quality management framework to be comprehensive including all relevant data quality dimensions. In order to assess the quality of risk modelling data the framework should include eight dimensions which are completeness, accuracy, consistency, timeliness, uniqueness, validity, traceability, and availability/accessibility. They underline that data quality standards should be set to all of these dimensions for all modelling input data and for each stage of the data life cycle. (European Central Bank 2018a)

In order to comply with the regulatory requirements, banking institutions should assess and measure data quality in an integrated and systematic way. The controlling activities need to cover the entire life cycle of data from entry to reporting and be applied for both historical and current data. The controlling activities need to be coherent among and across systems and include both internal and external data. The controls and procedures need to be planned for manual processes as well. The tolerance levels and thresholds should be clearly set for observing how the standards are met. Visual techniques are suggested for the representation of the indicators and quality levels set. (European Central Bank 2018a)

For data quality improvement purposes, banking institutions are given instructions to implement processes for identifying and overcoming quality deficiencies. An independent unit should undertake the assessment procedures. Based on the assessment, recommendations for correcting data with indication of priority should be given. The priority should be based on the materiality of the identified incidents. (European Central Bank 2018a)

When data quality deficiencies are found, all the incidents need to be recorded and monitored by the independent unit. Remediation plan should be formulated, and an owner appointed for resolving the issues. All the deficiencies need to be carefully resolved at source level rather than just mitigated. The schedule for the remediation is set based on the priority previously assigned and the time needed for implementation. (European Central Bank 2018a)

3.4 Required data quality dimensions

European Central Bank (2018a) states the data quality framework should assess the completeness of data. They define completeness as values being present in any attributes that require the information to be present (European Central Bank 2018a). Different researchers had a similar view on the definition of completeness dimension. Wang and Strong (1996) present a common understanding of completeness could be defined as the measure of broadness, depth, and scope of information hold within the data for its intended use. Ballou and Tayi (1998) define completeness as having all applicable information recorded. Olson (2003) talks about completeness under the accuracy dimension.

Batini and Scannapieca (2016b) state completeness means representing every relevant aspects of real world. Completeness is measured by comparing the content of the information available to the maximum possible content. (Batini and Scannapieca 2016b) In relational database context, Batini and Scannapieca (2016a) define completeness as the level of a table representing the real-life phenomena it is supposed to be representing. Completeness consists of the existence/lack and meaning of missing (NULL) values (Batini and Scannapieca 2016a).

The accuracy dimension has many elements. Researchers agree on the basic definition with ECB but also list different elements of what it means for data to be free of error. European Central Bank (2018a) requires data to be assessed by its accuracy. They define accuracy as data being substantively free of error (European Central Bank 2018a). Olson (2003) defines data accuracy as the measure whether values stored are correct and presented in a consistent and unambiguous form. Wang and Strong (1996) conclude accuracy being defined as the measure of data being correct, reliable, and provably free of error. Ballou and Tayi (1998) define accuracy as having correct facts representing the real-world event. Batini and Scannapieca (2016a) define accuracy as the closeness of the data value and the correct value aiming to represent the real-life event or object.

Batini and Scannapieca (2016a) divide accuracy to two definitions from the other one being structural accuracy and other temporal accuracy. Temporal accuracy refers to the rapidity with which the change in real-world object or event is displayed in the data value. Structural accuracy can be considered as syntactic accuracy and semantic accuracy. Syntactic accuracy checks whether a data value is part of the set of acceptable values. Semantic accuracy is defined as the closeness of data value to the true value. They argued that semantic accuracy is more complex to measure than syntactic accuracy. (Batini and Scannapieca 2016a p. 24)

The definition of consistency has conflicting views. Some researchers talk about consistencies across different sources and some across values. European Central Bank (2018a) states data should be assessed for consistency of data. They define consistency as any set of data matching across different data sources where the values represent the same events (European Central Bank 2018a). Wang and Strong (1996) conclude the definition of representational consistency as the measure of data being presented in the same format and being compatible with previous data. They also include describe consistency as data being consistently represented and formatted (Wang and Strong 1996). Batini and Scannapieca (2016a) define consistency as the semantic rules defined over data items not being violated. Semantic rules must be satisfied by all data values. They can be defined over an attribute or multiple attributes. (Batini and Scannapieca 2016a) Ballou and Tayi (1998) define consistency as the format being universal for recording the information. It could be concluded that the rules and formats should be consistent across different data sources.

European Central Bank (2018a) requires data to be assessed based on timeliness requirements. They define timeliness as data values being current and up-to-date (European Central Bank 2018a). Wang and Strong (1996) define timeliness as the measure of the age of data being appropriate for the task intended. Ballou and Tayi (1998) understand timeliness as having the information shortly after the real-world event. Batini and Scannapieca (2016a) talk about timeliness under the term of accuracy. They see timeliness as time-related accuracy dimension. Timeliness is defined as data being current and in time for their intended use. It is possible to have accurate and current data that is low-quality because of its uselessness since data is late for its intended use. Currency indicates data is being updated when the real-life events or objects are changing. High-quality timeliness dimension refers data being current but also available before its intended use. (Batini and Scannapieca 2016a pp. 27–28)

Uniqueness is not largely discussed in literature, but it is defined clearly by the regulators. European Central Bank (2018a) underlines that data should be assessed for uniqueness requirements. They define uniqueness as aggregate data not having any duplicate values arising from filters or transformation processes (European Central Bank 2018a). Batini and Scannapieca (2016a) talk about unique values under the accuracy measures. Accuracy can refer to sets of values as well, for example duplicate values when real-life object or event is stored more than once (Batini and Scannapieca 2016a). Wang and Strong (1996) mention uniqueness but do elaborate on its definition and scope. Uniqueness is thus assessed by the amount of duplicate values in this research.

Validity is not broadly discussed in literature. It is mostly mentioned under the term of accuracy. As validity is listed as one of the most important dimensions of data quality by the regulators, the dimension is discussed independent of accuracy in this research. European Central Bank (2018a) requires data to be valid. According to their definition, data validity means data is founded on a sufficient and thorough classification system that ensures their acceptability (European Central Bank 2018a). Batini and Scannapieca (2016a) name validity as part of accuracy. Olson (2003) agrees validity as part of accuracy dimension. Data validity means that a value should match one from the set of possible accurate values. Data validity does not necessarily mean it is accurate, since accuracy would also imply the value is correct. Defining the set of valid values for an attribute makes finding and rejecting invalid values relatively easy. (Olson 2003)

European Central Bank (2018a) states data should be available/accessible. They defined accessibility as data being available to all relevant stakeholders (European Central Bank 2018a). The definition of accessibility had similar definition in the literature. Batini and Scannapieca (2016a) define accessibility as the user’s ability to access information despite of culture, physical functions, or technologies available. For data to be accessible, data should be available or easily and quickly retrievable. Wang and Strong (1996) define accessibility as the level of data being available or easily and quickly retrievable. The role of IT systems is important for accessibility requirements to be met (Wang and Strong 1996). Assessing the accessibility of data is out of the scope of this study since it should include assessing the IT systems and the procedures of a company. Yet, it should be noted that accessibility is an important element in terms of the regulatory requirements.

Traceability is not widely discussed in literature but it is important in terms of the regulations. As the last dimension, European Central Bank (2018a) requires data traceability requirements to be met. They define traceability as being easily able to trace the history, processing practices, and location of the given data set. Wang and Strong (1996) concludes traceability being understood as the measure of how well data is documented, verifiable, and easily assigned to a source. Banking institutions should be able to trace the data back to its source systems and have the path well documented. As assessing the traceability would consists of assessing the IT systems and their information flows, it is not included in this research. It is still important to understand traceability as a major requirement to be complied with for the credit risk modelling data.

4 ASSESSING AND MEASURING DATA QUALITY

A literature review was chosen as the research method to get a comprehensive view on the methods that are generally used to assess and measure data quality. It was not necessary to identify all the possible researches and their evidence on the topic but rather discover different ways to address the issue of data quality assessment and combine different perspectives on it. This thesis is conducted for the purpose of identifying appropriate methodologies to assess and measure data quality for credit risk modelling purposes. Since all the methods presented are not from the banking field, the results only give indication on the possible assessment methods. The findings of the literature review could be used to improve the data quality management processes and metrics.

The research problem is first introduced by defining what the terms ‘assess’ and ‘measure’ mean. Assessing data quality means conducting a set of processes for the purpose of evaluating the condition of data. The aim of data quality assessment is to measure how well data represents the real-world objects and events it is supposed to be representing. The goal is to understand whether data meets the expectations and requirements for the intended use. Measuring is essential for comparing different objects across time. For effectively measuring data quality, measurements should be comprehensible, interpretable, reproducible and purposeful. The context should be understood for interpreting the measurements and it should be clearly defined what the measurements represent and why they are conducted. For comparing the improvement or deterioration of measures, it is necessary to be able to repeat the measurements the same way over time. (Sebastian-Coleman 2013 pp. 41–47) In this thesis, the subchapters are divided into assessment techniques and measurement techniques. The assessment techniques list all the methods that could be used as indication on the level of data quality. All the techniques presented should be adjusted for the specific purpose of the data. The measurement techniques present the precise metrics and formulas how the level of data quality could be quantified or presented.

The process of the literature review methodology is presented in figure 2. Relevant literature was searched from university’s library sources, Web of Science and Scopus. There were only few papers available thus enough papers from the financial field could not be found and the searches were done generally to all fields. The number of articles found was large using searches without specifying the field thus the searches were limited by searching the key words from either title or abstract. The following key words were used in different combination: “data quality” or “information quality”, assess* or measur* or valid* or examin*, completeness or accuracy or consistency or timeliness or uniqueness or validity. The search was limited to articles published in English language, and they needed to be publicly available or available using university’s account . The titles and the abstracts of the first 200 papers ordered by the search relevance from four different searches were read through and the most relevant were selected. Based on the title and abstract, 84 studies were included. Finally, the selected studies were read through and the irrelevant ones were removed. The removed articles were either duplicates, or they concentrated on the general issues in data quality or the issue of choosing the right dimensions rather than measuring data quality. The analysis then consisted of 39 papers which were read through and further analyzed. After the selection of suitable articles, their references were scanned through to find more studies on the subject. Three additional items were found. Also, two books were hand-picked as they were found relevant while conducting the theoretical framework of this thesis. The literature review finally consisted of 44 sources.

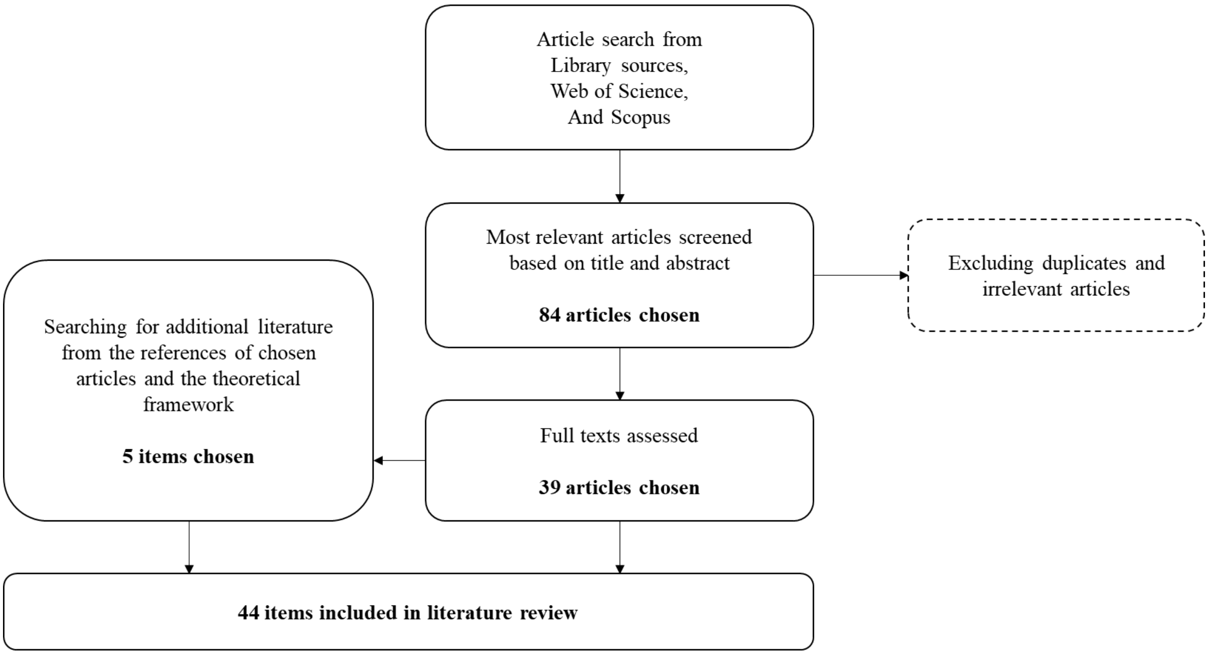
Article search from Library sources, Web of Science, And Scopus
Most relevant articles screened based on title and abstract
84 articles chosen
Excluding duplicates and irrelevant articles
Searching for additional literature from the references of chosen articles and the theoretical framework
5 items chosen
Full texts assessed
39 articles chosen
44 items included in literature review

Figure 2. The phases conducted in the literature review process

While going through the literature review results, it occurred that some of the articles were trying to examine a specific database while defining the dimensions and metrics used, while others were with the objective to find appropriate measures or measuring procedures in general. To get a holistic view on the subject, both types of sources where considered. The articles and books chosen for the literature review are presented in table 1. The dimensions that were mentioned in the article are listed in the table. If the authors used a similar definition to that of the selected six dimensions but did not name the dimension with the same term, it was not included unless one of the six dimensions was named as a synonym. For example, Heinrich and Klier (2011) said "Often [currency] is also referred to as timeliness and sometimes it is even seen as a part of timeliness" thus it was included as timeliness dimension. In some literature, some of dimensions were discussed under other dimensions. Olson (2003), for example, discussed completeness and validity under the term of accuracy. In these cases, only the hypernym discussed was marked to the summary table (table 1) but the results are discussed under the different dimensions. The articles could have discussed other dimensions as well: the dimensions of concordance (Akhwale et al. 2018), comparability (Lambe et al. 2015; Heikkinen et al. 2017; Arboe et al 2016; Bah et al. 2013; Asterkvist et al. 2019; Bray et al. 2018; Jonasson et al. 2012), redundancy (Chen et al. 2015), sparcity (Chen et al. 2015), reliability (Blevins et al. 2012; Weidema and Wesnaes 1996), correctness (Liaw et al. 2015), and representativeness (Lim et al. 2018) were included in the reviewed papers. Only the dimensions of the scope of this study are included in the summary. If the research did not mention dimensions at all but discussed data quality generally, it was included in table 1 with a dimension column of "Not specified". The field of the study was summarized so that the whole article was concentrating to that field or the authors used a sample data set from that field. If the article did not concentrate on a specific field or had examples on several fields, the field was marked as "not specified".

Table 1. The sources used in the literature review

Source	Field	Completeness	Accuracy	Consistency	Timeliness	Uniqueness	Validity	Not specified
Abela et al. (2014)	Healthcare							■
Akhwale et al. (2018)	Healthcare	■						
Amoroso et al. (2014)	Healthcare	■		■				
Anderka et al. (2015)	Healthcare	■	■		■			
Arboe et al. (2016)	Healthcare	■					■	
Arts et al. (2002)	Healthcare	■	■					
Asterkvist et al. (2019)	Healthcare	■			■		■	
Ayatollahi et al. (2019)	Healthcare	■	■	■	■			
Bah et al. (2013)	Healthcare	■					■	
Baesens et al. (2010)	Banking	■	■	■	■			
Barker et al. (2012)	Healthcare	■	■					
Blevins et al. (2012)	Healthcare		■					
Borek et al. (2011)	Not specified							■
Borek et al. (2013)	Not specified	■	■					
Box et al. (2012)	Healthcare	■			■		■	
Bray et al. (2009)	Healthcare	■			■		■	
Bray et al. (2018)	Healthcare	■			■		■	
Bray and Parkin (2009a)	Healthcare				■		■	
Bray and Parkin (2009b)	Healthcare	■						
Busacker et al. (2017)	Healthcare	■			■			
Charrondiere et al. (2016)	Nutrition							■
Chen et al. (2015)	Not specified		■					
Clayton et al. (2013)	Healthcare		■					
Crocetti et al. (2001)	Healthcare	■						
Espetvedt et al. (2013)	Dairy	■						
Ezell et al. (2014)	Aviation	■	■	■				
Fisher et al. (2009)	Not specified		■					
Gray et al. (2015)	Healthcare	■		■			■	
Habibi et al. (2016)	Trade	■	■	■				
Heikkinen et al. (2017)	Healthcare	■			■		■	
Heinrich and Klier (2011)	Telecommunication				■			
Hinterberger et al. (2016)	Nutrition							■
Holden (1996)	Nutrition		■					
Jonasson et al. (2012)	Healthcare	■			■		■	
Lambe et al. (2015)	Healthcare	■			■		■	
Lee et al. (2006)	Not specified	■	■	■	■			
Li et al. (2014)	Not specified		■					
Liaw et al. (2015)	Healthcare	■		■				
Lim et al. (2018)	Healthcare	■	■		■			
Liu et al. (2014)	Banking	■	■	■	■	■	■	
Majumdar et al. (2014)	Aviation							■
Olson (2003)	Not specified		■					
Sadiq et al. (2014)	Retail	■	■	■	■			
Weidema and Wesnaes (1996)	Not specified	■						

From the selected 44 items, 2 (4.5 %) were from the banking field, and 25 (56.8 %) were from the field of healthcare. In 8 (18.2 %) articles the field was not specified, and in 9 (20.5 %) the field was other than those mentioned. Many of the articles from the healthcare field were researching the data quality of cancer registries. As it was previously discussed in this thesis, quality depends on the use case. As the medical field deals with human lives and safety, it requires a high degree of quality. Many studies in the medical field assessed the data quality of population-based cancer registries. Those registries are highly significant in estimation of cancer survival thus they require strict data quality controls (Abela et al. 2014). The studies in other fields also often included the safety of humans (such as aviation safety) which requires high degree of quality. Thus, similar techniques to assess data quality could be applied in the banking industry as the needed level of quality is not notably higher than in the fields reviewed.

Different terms were used in the literature to represent data records or attributes. In this thesis, the terms record and attribute are used. The term record refers to set of data values that represent tuples or table rows in a relational database. The term attribute is used to represent data fields or table columns.

4.1 The overall quality of data

In some papers, the authors described data quality controlling methods that were not specifically intended to a certain dimension or the author did not cover data quality dimensions at all. These measuring methods could give ideas or indication what could be examined to detect data quality issues of certain dimension. The methods without a mention on certain dimension are presented first so that they could possibly be matched with a certain dimension based on other findings from the literature.

4.1.1 Assessing the overall quality of data

Abela et al. (2014) studied data quality controls of population-based cancer registries in order to estimate cancer survival. The authors used the terms of data being valid, accurate, comparable or complete but they did not name them as data quality dimensions or present tests under a specific dimension. They presented a method which consisted of three phases. They assessed data by each value but also as a whole set. Their method was mostly based on rules defined by professionals. First phase assessed the attributes independently, second phase assessed individual records and the third phase the whole data set. During the first phase, individual attributes were assessed based on whether they obeyed the protocols. The type of data (for example numeric), number of digits or characters and valid values were recorded for each attribute. Also, it was examined what value is used if the value is missing or if it is possible to be missing. Each value needed to fit within the specific range or otherwise meet the definitions for that specific attribute. They stated values containing errors should be examined and corrected, and documented. (Abela et al. 2014) The selected attributes and the rules for correct values are presented in table 2.

Table 2. Validity rules for researched attributes in a cancer registry presented by Abela et al. (2014)

Attribute	Type	No. of digits/characters	Valid values	Value used when missing
Unique ID	Alphanumeric	Depends on the source		Not allowed
Sex	Numeric	1	1,2	9
Day of birth	Numeric	1 or 2	1-31	99
Month of birth	Numeric	1 or 2	1-12	99
Year of birth	Numeric	4	Depends on the scope of analysis	9999
Day of diagnosis	Numeric	1 or 2	1-31	99
Month of diagnosis	Numeric	1 or 2	1-12	99
Year of diagnosis	Numeric	4	Depends on the scope of analysis	9999
Last known vital status	Numeric	1	1-3	9
Day of last known vital status	Numeric	1 or 2	1-12	99
Month of last known vital status	Numeric	1 or 2	1-12	99
Year of last known vital status	Numeric	4	Depends on the scope of analysis	9999
ICD-O-3 topography	Alphanumeric	4	C00.0-C80.9	Not allowed
ICD-O-3 morphology	Numeric	4	8000-9989	9999
Behavior	Numeric	1	0,1,2,3,6,9	Not allowed

Secondly, every record was checked and excluded if ineligible or incoherent. This included examining e.g. the coherence of dates, missing values, the registration type, and duplicate registrations. The criteria were defined as logical rules in their use-case such as if the behavior code was 2 the record was ineligible for the analysis, or if the age was more than 100 the record was excluded from the analysis. The rules were chosen so that a specific cancer estimation would be accurate, consistent and comparable. The records failing one or more criteria were flagged and further analyzed and revised. Finally, the distribution of key characteristics of the dataset was assessed. The last phase assessed the whole data set for record proportions, distributions, counts over time etc. The techniques included calculating the proportion of death-certificate only (DCO) registrations over time, the proportion of tumors morphologically verified, and the distribution of cancers by population and age over time. The results of the third phase should be accessible for the users for the ability to compare between systems. During the comparison, it is important to notice that the differences of proportions could be caused by changes in coding practices. For example, tumors that were previously classified as invasive were later excluded from the data. (Abela et al. 2014) The changes of coding practices should be analyzed and documented. If the changes are not taken into account, inconsistencies occur in the analysis.

Hinterberger et al. (2016) researched data quality requirements for establishing a framework for managing food composition data. A system for food composition data is important for policy makers and researchers. They argued a data quality assessment based on questions made by EuroFIR project was not enough since values with large errors could get a high-quality index. The EuroFIR project questions were based on food description, component identification, sampling plan, number of analytical samples, sample handling, analytical method and analytical quality control. Each of these were scored from 1 to 5 points depending on how many ‘yes’ answers they contained and then summed up to reach the final score. (Hinterberger et al. 2016)

To improve the data quality assessment Hinterberg et al. (2016) first collected data quality requirements in the field of food composition. First, the basic information on the attributes were collected such as data type, whether it’s a primary key, whether it is mandatory or not, and whether it’s a set-value. The attribute details expose certain restrictions on the database system. Primary keys uniquely identify the data records and it seen as a mandatory attribute for ensuring the functionality of database tables, characters cannot be typed when number is expected and set-values need to belong to a fixed set of values. The requirements were collected through entity details documentation, quality index guidelines documentation, and domain experts. They finally defined 451 data quality requirements from which 329 from different guideline documents and 122 from logical reasoning and domain experts. The requirements included rules for independent attributes and also rules that depend on several attributes. A list of the requirements was not provided. They scored all the requirements and found three groups: hard constraints, soft constraints, and indicators. Hard constraints are significant in order to understand data and they always indicate invalid data thus they should be checked when data is entered. Soft constraints affect quality to some extent but do not make data invalid (for example, not enough significant digits or wrong classification). Indicators have least contribution to data quality. If there appears to be deficiencies in the data but it’s not certain, it is an indicator (for example, data component has been changed a little bit over time). They talked about preventing the quality issues (rules and instructions while entering values). However, preventing is not always possible. (Hinterberger et al. 2016)

Charrondiere et al. (2016) presented a table of proposed checks presented by Food and Agriculture Organization (FAO). The checks were categorized into four themes: food identification, component checks, recipe checks, and data documentation. The checks were specific to food composition data and each check was specifically defined for particular food composition issue. The tests included business rules for specific values, mathematical checks, comparability checks, systematic checks, missing values checks, documentation checks, and processing method checks. The examples included tests on consistency of value naming and the use of singular and plural values, duplicate values, values complying with rules assigned, minimum and maximum values in a specific range, standard deviation calculation, external values converted correctly, the sum of values within acceptable range, the consistency on definitions/formulas used, correct language, comprehensive documentation, missing values, the source and calculating methods included, order sorted within a specific group. (Charrondiere et al. 2016)

Borek et al. (2011) listed data quality assessment methods based on literature and expert knowledge on current practice. They divided the methods into nine categories. These included attribute analysis, cross-domain analysis, data validation, domain analysis, lexical analysis, matching algorithms, primary key and foreign key analysis, schema matching, and semantic profiling. (Borek et al. 2011) The techniques are summarized in table 3.

Table 3. Data quality assessment methods based on Borek et al. (2011)

Method	Description
Attribute analysis	Calculating and assessing: - Number of values - Number of unique values - Number of instances per value as percentage from the total - Number of NULL values - Minimal and maximal value - Total value - Standard deviation - Median and average value - Inferred type information - Frequency distribution - Format distribution
Cross-domain analysis (Functional dependency analysis)	Comparing the percentage of values within attributes across attributes from different tables
Data validation	Verifying values against a reference data set - in manual validation, a sample is selected - in automated validation, a complete dataset is validated
Domain analysis	Verifying if data values are within - a specific series of values - a predefined set of values - predefined range conditions
Lexical analysis	Mapping unstructured content to a structured set of attributes by rule-based or supervised-model based techniques such as phonetic algorithms
Matching algorithms (Record-linkage algorithms)	Identifying duplicate records
Primary key and foreign key analysis	Analyzing whether an attribute could be included in the primary key/foreign key relationship
Schema matching	Using database schema matching algorithms to detect whether two attributes are semantically equivalent
Semantic profiling	Verifying data against specified business rules

Majumdar et al. (2014) proposed a framework to assess the quality of external data. Their methodology was based on Multi-Criteria Decision Analysis (MCDA). The MCDA follows an eight-step process which consisted of defining the objective, identifying options, identifying the option assessing criteria, scoring the options, identifying the weighting of criteria, aggregating the scores and weights, assessing the results, and finally conducting sensitivity analysis. They examined 12 different aviation safety databases. The assessment criteria were based on the authors' understanding of reporting criteria and expert-validated documentation made by International Civil Aviation Organization. The authors then compared the values in the databases against predefined criteria. (Majumdar et al. 2014) The criteria and the possible outcomes are presented in table 4. The formulas for the scoring of the database are later presented in measuring techniques (see 4.1.2).

Table 4. Assessment criteria for aviation safety databases presented by Majumdar et al. (2014)

Criteria	Possible outcomes
Time of report	On the day of occurrence Later Unknown
Major form of reporting	Electronic reporting system Email, post, telephone Mixed of electronic reporting system and email, post, telephone
Level of investigation	All occurrences Most occurrences (>20/year) Less than 20/year (only high-severity occurrences) None
Bias	From 0 to 7 types of bias
Database translation required	Yes/No
Feedback given to reporter	Yes/No
Publication of statistics/safety report	From 0 to 5 types of publications
Data sharing with other stakeholders apart from regulatory requirements	Yes/No
Data accessibility	Original report Primary database Report written by the primary organization Secondary database Tertiary database
Source of descriptive narrative	Reporter + primary investigator Reporter + primary analyst Reporter + secondary investigator Reporter + secondary analyst Primary investigator Primary analyst Secondary investigator Secondary analyst No narrative
Consistency of the reporting system over time	Consistent/Inconsistent

Funk et al. (2006) discussed data quality could be assessed by applying a data quality survey for data collectors, data managers, and data consumers. The survey reflects the respondent's view on different data quality dimensions. The responses give a score from 0 to 10 for different statements such as "this information is of sufficient volume for our needs". The results should then be analyzed to draw conclusions on the quality. (Funk et al. 2006 pp. 31-34) A questionnaire to different stakeholders was conducted by Ayatollahi et al. (2019). Also Baesens et al. (2010) assessed data quality in the banking field. They used a targeted questionnaire that consisted of 65 questions of data flow from source to end (Baesens et al. 2010). This method is subjective and does not present the quality objectively. Funk et al. (2006) argued a good assessment approach is to conduct a comparative approach which combines subjective and objective assessment. The comparative approach is done by performing data quality survey and using data quality metrics, then by comparing the results and analyzing the root-causes for differences. That way, data quality can be given quantitative metrics but the knowledge of stakeholders is also taken into account. The quality survey additionally raises awareness of the issue inside the organization. (Funk et al. 2006)

4.1.2 Measuring the overall quality of data

Hinterberger et al. (2016) stated the most important requirements for a data quality framework is to provide clear visual access to data quality information and metrics to understand both an overview and ability to select a detailed view. The data quality analysis should also provide the user information on the actions to be taken to address quality issues. (Hinterberger et al. 2016) The simplest method to quantify data quality gives a binary value to present whether the attribute value fulfills the predefined criteria of quality or not. Hinterberger et al. (2016) suggest assessing each value for each data quality requirement and then giving it a value of 1 if it fulfils the requirement and 0 if not, then calculating the average for all the data points. (Hinterberger et al. 2016) Funk et al. (2006 p. 54) present the data quality assessment measure as

$$M_i = 1 - \frac{Number\ of\ undesirable\ outcomes_i}{Total\ outcomes_i} \tag{2}$$

where 1 represent the optimal score and 0 the poorest outcome. After the rating is calculated for each attribute i, the results can be aggregated for all records by using minimum or maximum operator. For example, minimum assigns a quality level of the weakest quality attribute. Alternatively, a weighted average of the attributes could be calculated. The formula is given as

$$M_{agg} = \sum_{i=1}^{card(T)} g_i M_i , \quad where\ g_i \epsilon [0; 1]\ and\ 0 < g_i < 1 \tag{3}$$

where g_i is a weighting factor and M_i is a value of the assessment of the *i*th attribute for a specific dimension, *T* is the set of attributes and *card(T)* the number of attributes. (Funk et al. 2006 p. 57) Majumbar et al. (2014) had predefined criteria for data quality which were scored from 0 to 1. The possible outcomes for each criterion were identified and the scoring was decided so that the most ideal option was given a value of 1 and the least ideal a value of 0, the rest everything in between based on their importance. The criteria were assigned equal weights but they could have been given weights based on their importance to the decision. Finally, the Weighted Sum Method (WSM) was used to calculate the overall score of the database. The score was presented as

$$S_i = \sum_{j=1}^{n} w_j x_{ij} \quad i = 1,2, \dots , m \tag{4}$$

where x_{ij} is the scoring of the *i*th option on the *j*th criterion, and w_j the weight of the *j*th criterion. In their study, 11 criteria were assessed with equal weights thus the ideal system would have an overall score of 11. They compared the scores of different databases so the overall score of each database was normalized. The normalized scores were obtained by

$$S_{i\ normalized} = \frac{S_i - S_{MIN}}{S_{MAX} - S_{MIN}} \tag{5}$$

where S_{MIN} present the minimum score and S_{MAX} the maximum score. The authors then presented the results of each reporting system in relation to an ideal system. The sensitivity analysis could contain the scoring of the options and the weights assigned. (Majumdar et al. 2014) By this method, data quality of a database could be given a score and compared to an ideal database.

4.2 Completeness of data

The completeness dimension covers the assessment of NULL values. Inaccuracies occur when objects are missing or have missing pieces. (Olson 2003) Missing value or record generally means an event or object exists in real life but for some reason is unavailable in data. There are three different cases for NULL values to occur. First, it might be that the value exists, but it is not known. Secondly, the value does not exist at all. Or lastly, the value might exist but is not known whether it should exist or not. (Batini and Scannapieca 2016a) In the case of credit risk modelling data, completeness can sometimes be a problematic concept. When it is mandatory for an attribute to have a value, errors caused by missing values can be easily detected but when an attribute may or may not have a value, the assessment is not as straightforward.

Missing values and missing records can be very problematic to identify. A database could be correct on every level except that missing information causes quality problems. Missing data value may or may not be correct. For an attribute that always should have a value, missing values are incorrect. Olson sees missing values as particularly problematic when a missing value could be sometimes correct and sometimes not. (Olson 2013) In a database, both a record that has an attribute with no value and missing values are represented as NULL values. A NULL value is correct when the event does not exist in real life, such as person not having default – thus default date should not have a value. The value would be incorrect if a person had default but the default date is not filled. Baesens et al. (2010) refers to these as causal and not causal missing values. Causal missing values are accepted missing values which indicate there is an acceptable reason for missing values to occur. For example, the attribute containing information on the tax number should not have a value when the client is a private person. (Baesens et al. 2010)

Inaccuracies of data are easily occurring if a system does not make distinction between a correct missing value and a value that was not filled because the correct value was not known. The issue of problematic NULL values should be addresses by creating a new attribute that indicates whether the value of the other attribute is missing or not. Creating a value for the attribute that indicates missing value, such as “not part of an event“, would be bad practice since then data aggregation does not formulate correctly. In business context, these NULL indicator attributes are rarely used. Even if they would be used, all the users might not use them correctly which would again cause inconsistencies. (Olson 2003) Since there are no best practices for preventing missing data, some metrics are needed for measuring the completeness of a database.

Assessing the completeness of data was included in most of the research papers. Thus, it could be argued that it is one of the most important components of data quality in the fields present in the search. The papers addressed the completeness dimension in two ways: either internally measuring the extent of values missing from an attribute or record where they should be present, or externally assessing the extent of records that were not stored at all. Some studies examined the completeness by using multiple methods and some by only one. Some calculated both the completeness of attributes or records and the completeness of whole data set, others concentrated in either one. The completeness was is some studies measured by using a randomly selected sample and then estimated to the whole dataset, and some examined the whole dataset. The subchapter is divided to two. First, all the assessment methods are presented. Then, the thesis covers how completeness could be measured and quantified.

Some researchers divided the measuring methods into two categories: semi-quantitative and quantitative. This division was done by Bray and Parkin (2009b) and Heikkinen et al. (2017). Bray and Parkin (2009b) defined semi-quantitative methods as techniques to gain insight of the degree of missing records over time or missing records compared to another database. They defined quantitative methods as techniques to assess the extent of missing records numerically (Bray and Parkin 2009b). As they had been researching the techniques to measure completeness in the medical field, their research provided the definitions for many of the metrics introduced. Both semi-quantitative and quantitative methods are described under the assessment techniques. The more detailed metrics for quantitative methods are presented under the measuring techniques.

4.2.1 Assessment techniques

Most studies assessed the completeness internally by simply examining the number of missing values (Akhwale et al. 2018; Amoroso et al. 2014; Barker et al. 2012; Borek et al. 2013; Ezell et al. 2014; Gray et al. 2015; Habibi et al. 2016; Liaw et al. 2015; Lim et al. 2018; Sadiq et al. 2014). The number of missing (NULL) values can be calculated for each attribute separately. The extent of missing values can be calculated for the whole data set or for a small data sample. Akhwale et al. (2018) first defined the attributes that should contain values and then calculated the proportion of missing or invalid values for any of them. Amoroso et al. (2014) examined completeness for 10 indicators selected as top priority.

Some researches defined different levels of performance and assessed the completeness against those levels (Anderka et al. 2015; Weidema and Wesnaes 1996). Anderka et al. (2015) examined data quality measures for population-based birth defects surveillance. They evaluated the extent of birth defects, the extent of different pregnancy outcomes included and the extent of data elements collected for different birth defect programs. They defined the optimal, the essential and the rudimentary level to each separately. For example, for pregnancy outcomes: the rudimentary quality included only live births, the essential level live births and miscarriages, and the optimal level consisted of all live births, miscarriages, and other pregnancy losses. (Anderka et al. 2015) The different levels of quality used by Anderka et al. (2015) are presented in table 5. Completeness is thus examined by defining what a database should elementarily contain and optimally contain or anything between, and finally assessing the situation against those definitions.

Olson (2003) discussesses completeness under the term of accuracy. Olson names value rule analysis as a possible method to reveal uncomplete data. Value rule analysis is used for trying to find unreasonable results through data aggregation. The analysis can be done by using cardinality, counts, sums, averages, medians, frequency distributions, standard deviations and other similar aggregations. Any aggregation test that allows the analyst to investigate values for completeness or reasonability can be used. Value rule analysis does not provide excessive report of the values but rather an overview of the reasonability. Extreme results can be easily detected but the values between reasonable and unreasonable need some extra inspection. Basic rules should be collected from data users to understand what results are expected before executing the tests. The rules could include for example the expected range of percentage of a specific value or the expected frequency of a specific value. When executing value rule analysis, all the data should be investigated over a certain period of time. Finally, the results should be validated with a group of field specialists. If needed, the tests and expected results can be modified based on new information revealed from the results. (Olson 2003) Examining aggregated value frequencies had specific techniques in the medical field such as historical trends technique, mortality:incidence ratio technique, and histological verification technique.

Assessing historical trends was named as one of the semi-quantitative techniques to give indication on the completeness of data by Bray and Parkin (2009b). The objective was to detect unexpected or improbable trends, or to compare the obtained trends from the data to the results gained from another source or population. If the trends are compared to another source, a “gold standard” needs to be defined and statistically significant differences flagged. (Bray et Parkin 2009b) In the medical field, a common technique was to assess incidence rates (Bah et al. 2013; Bray et al. 2009; Bray et al. 2018; Heikkinen et al. 2017; Jonasson et al. 2012). Bray et al. (2009) and Bray et al. (2018) also studied the stability of data over time. Bray et al. (2018) analyzed the stability of incidence rates and compared them to other countries from the region. Finally, they presented the trends by drawing age-specific graphs and analyzed the shape of them. The graphs are used to show the incidence rates across different ages. The expected shape of the curves were based on biological characteristics of the diagnosis. A decrease in the curve could be due to missing cases. For example, a decrease in the incidence rates of older people might indicate that elderly patients are not as accurately diagnosed. (Bray et al. 2018)

Another semi-quantitative method described by Bray and Parkin (2009b) was the mortality:incidence (M:I) ratios which represent the relationship between the number of deaths and the number of new incidents. This method was used in many papers to assess completeness. The ratios were first calculated and then compared with the region standards using a predefined significance test. The mortality:incidence ratios were calculated by Bray et al. (2009), Bray et al. (2018), Heikkinen et al. (2017) and Jonasson et al. (2012). Bray et al. (2018) fitted a regression line to the survival rates and examined the deviations. Also one semi-quantitative method named by Bray and Parkin (2009b) was the histological verification of diagnosis. The percentage of cases morphologically verified (MV%) was used in the medical field because histological verification is seen as a reliable method for incidence validation. The percentage was then compared to expected values. (Bray and Parkin 2009b)

Bray and Parkin (2009b) named independent case ascertainment as the first quantitative method. It included two methods the first methods being case-finding audits. They explained the idea of audits is to identify the cancer cases from original data such as medical records during a certain period of time and detect the missing cases that were not recorded on the database. (Bray and Parkin 2009b) In some papers, the completeness was assessed by audits (Anderka et al. 2015; Arboe et al. 2016; Box et al. 2013; Arts et al. 2002). Box et al. (2013) calculated the percentage of key attributes that were included in the laboratory report.

The second independent case ascertainment method was comparison of two or several independent sources (Bray and Parkin 2009b). Bray and Parkin (2009b) argued the comparison of two or more independent sources was a good technique to assess the completeness. The comparison is done by linking records between databases and calculating the number of cases that are missing from one database even though they are registered in another database (Bray and Parkin 2009b). The completeness was measured by comparing to another source by Asterkvist et al. (2019), Busacker et al. (2017), Espetvedt et al. (2013), Heikkinen et al. (2017), Jonasson et al. (2012), Arts et al. (2002) and Lambe et al. (2015). Asterkvist et al. (2019) and Lambe et al. (2015) had as a “gold standard” a cancer registry to which registration was mandatory.

Another quantitative method was the capture-recapture method (Bray and Parkin 2009b). The capture-recapture method were mentioned as a completeness measure by Bah et al. (2013), Bray et al. (2009) and Crocetti et al. (2001). Crocetti et al. (2001) applied the capture-recapture method to estimate the completeness of a population-based cancer registry. The capture-recapture method assumes that every capture is independent of each other. In other words, the independency implies that an incident being recorded in one source is not modified by the result of an incident being recorded in another source. (Crocetti et al. 2001) Brenner (1995) states negative dependence between the sources lead to underestimation of completeness and positive dependence to overestimation of completeness. Crocetti et al. (2001) named three sources from which a cancer incident can be recorded from: clinical, pathological and death certificate. Clinical registration of cancer is based on information from hospital discharge or general practitioners, pathological from histological reports and death certificate from death certificates where tumor is reported as death cause. (Crocetti et al. 2001) Bah et al. (2013) argued in the case of cancer registries, the sources are unlikely to be independent. Thus two sources of the most dependence were grouped together and the number of missing cases was examined against the third source. (Bah et al. 2013) Similarly Crocetti et al. (2011) grouped the two most dependent sources together and used the capture-recapture method between the grouped source and the third source.

The last quantitative method defined by Bray and Parkin (2009b) was the death certificate (DC) method where completeness was calculated by the proportion of registered incidents that were registered via a death certificate. The methods could be used for the whole data or to subsets of it. The DC & M:I method is used to have an approximate measure of how many unregistered cases did not die by assuming the proportion of unregistered cases that caused death is the same as the proportion of registered cases that caused death. Another method was the Flow method which similarly calculates the approximation of the incidents not traced via death certificates. The objective is to find the unregistered incidents which did not cause death or the incident that caused death but the real cause of death was not mentioned in the death certificate. The formula for the method is presented in the measurement techniques chapter. (Bray and Parkin 2009b) Bray et al. (2009) used the Flow method as one of the techniques to assess the completeness of a national cancer registry.

The last semi-quantitative method named in the field of healthcare by Bray and Parkin (2009b) was assessing the number of sources. In order to get indication of the completeness, the average number of sources and the average number of notifications per case could be calculated. In the medical field when multiple sources are used to check for an incident, the probability of case being unreported decreases. (Bray and Parkin 2009b) The number of sources or incidents was used as a completeness measure by Anderka et al. (2015), Bah et al. (2013), Bray et al. (2009), Heikkinen et al. (2017) and Jonasson et al. (2012). Anderka et al. (2015) calculated the number of systematically used sources and assessed it against three levels of performance defined beforehand. The guidelines of the National Birth Defects Prevention Network (NBDPN) provided the basis for the chosen measures (Anderka et al. 2015). The different quality levels for sources used (DQ1.1) are presented in table 5.

Table 5. Completeness levels of a birth defect database presented by Anderka et al. (2015)

Performance measure	Level 1 Rudimentary	Level 2 Essential	Level 3 Optimal
DQ1.1 Types of data sources used systematically and routinely to identify potential cases at a population-based level	Each of the following sources: vital record data, additional source for case identification	The data sources in level 1 and any additional sources of natal or postnatal data	The data sources in levels 1 and 2, as well as routine reporting from any of the following data sources b for systematic specialized ascertainment of prenatally diagnosed defects
DQ1.2 Birth defects included using standard NBDPN case definitions	All of the NBDPN “core” birth defects	All of the NBDPN “recommended birth defects”	Major structural malformations beyond those birth defects identified on the NBDPN list
DQ1.3 Pregnancy outcomes included	Live births	Live births, stillbirths	Live births, stillbirths, and other pregnancy loss
DQ1.4 Systematic and routine identification of cases (age of diagnosis) during ascertainment period	Identification of cases diagnosed through 1 month of age	Identification of cases diagnosed through 1 year of age	Identification of cases diagnosed beyond 1 year of age
DQ1.5 Data elements collected	All “core” data elements	All “recommended” data elements	All “enhanced” data elements

All the techniques found in this thesis were divided into six method categories. A summary of the methods used for assessing completeness is presented in table 6. The table describes what techniques belong to each of the methods.

Table 6. Summary of the reviewed methods to assess the completeness of data

Method	Description
Number of NULL values	Calculating the amount of missing information
Rules for different levels of completeness	Defining the desired level of completeness based on what the records should contain and comparing the situation to the predefined levels
Value rule analysis	Searching for unreasonable results through data aggregation, assessing historical trends
Case-finding audit	Recoding original information and comparing the database to the recoded values
Comparing sources	Linking records between two or several databases and identifying the values that exits in all of them, only some of them or none of them Methods such as capture-recapture method, DC&M:I method, Flow method
Source analysis	Analyzing the reliability of a source and the average number of sources where a case is identified

Liu et al. (2014) assessed data quality in the banking industry. They argued financial institutions should assess completeness of interface files in transmission and record counts with a formula to get a numerical estimation. (Liu et al. 2014) They did not present any formula in their study. For this purpose, formulas found in the reviewed articles are presented next.

4.2.2 Measuring techniques

Olson (2013) reminds the quality measures of completeness depend on the intended use. Data might be good-quality for some use, but the same data could be poor-quality for another use (Olson 2003). In the reviewed articles, the measured completeness rates were analyzed based on the needs of a specific case. Between the papers, no consensus existed on what should be the exact level of high-quality.

The basic metrics for completeness was to measure the extent of values missing for an attribute or data record (Akhwale et al. 2018; Amoroso et al. 2014; Barker et al. 2012; Borek et al. 2013; Ezell et al. 2014; Gray et al. 2015; Habibi et al. 2016; Liaw et al. 2015; Lim et al. 2018; Sadiq et al. 2014). Funk et al. (2006 p. 56) present completeness metric as a simple measure defined as

$$Completeness_i = 1 - \frac{Number\ of\ incomplete\ values_i}{Total\ number\ of\ values_i} \tag{6}$$

Akhwale et al. (2018) created binary value to present whether any of the attribute value was missing, and then calculated the proportion of missing or invalid values for any of them. They only had a sample of randomly selected records. By using a generalized estimating equation model with a log link, binomial distribution, exchangeable correlation matrix, and robust standard errors they assessed the total risk of having missing values. (Akhwale et al. 2018) The formula was not presented in their research. Also Ezell et al. (2014) created a binary value for presenting whether the record was complete or not. They measured completeness of all the 14 attributes that they had selected in their study. The completeness was defined as

$$IC_{ij} = \begin{cases} 0 & if\ the\ value\ is\ complete \\ 1 & if\ the\ value\ is\ incomplete \end{cases} \tag{7}$$

for *i=1, ..., 14* attributes within *j=1, ..., N_R* part records. They then estimated the proportion of complete values from a sample data. For 11 attributes, no incomplete values could be found from the sample. When the probability of finding incomplete values is small, a large sample is needed to find one. (Ezell et al. 2014) When incomplete values were not found, Ezell et al. (2014) used a Bayes estimator proposed by Zhang et al. (2013). The Bayes estimator was calculated as

$$\hat{p}_{OB} = \frac{N + a}{m + a + b} \tag{8}$$

where *N* presents the number of incomplete values in the sample of m records. a and b presents the prior numbers of values that are incomplete and complete (Zhang et al. 2013). Zhang et al. (2013) suggested using *a*=1 and *b*=999 when the percentage of defects observed was less than the true percentage. Ezell et al. (2014) used the suggested values for a and b when incomplete values were not found from the sample. If incomplete values were found, they calculated the maximum likelihood estimates. The maximum likelihood estimator was calculated as

$$\hat{p}_O = \frac{N}{m} \tag{9}$$

where *N* presents the number of incomplete values and *m* the total number of records. (Ezell et al. 2014) Amoroso et al. (2014) calculated completeness as the proportion of the number of values not missing. They calculated the completeness rate as the sum of non-missing values across all 10 indicators divided by the expected number. In additional to calculating the completeness rate for each indicator, it was also calculated to the whole data set of each district. The total reporting completeness was calculated by comparing the number of monthly reports received to the expected number. (Amoroso et al. 2014)

Some researchers defined different levels of performance and assessed the completeness against those criteria (Anderka et al. 2015; Weidema and Wesnaes 1996). Anderka et al. (2015) defined the optimal, the essential and the rudimentary level to each separately. The different levels could be each given a quality score. Weidema and Wesnaes (1996) also gave the completeness dimension a score from a scale from 1 to 5. They defined what does each score mean for completeness indicator. (Weidema and Wesnaes 1996)

When comparing the incidence rates of several sources, the incidence rates could differ vastly or only by a small decimal. Thus, it should be defined what is accepted as the same value. Bah et al. (2013) used statistical methods to calculate the similarity between the rates from two sources. If completeness was assessed by comparing the database with the original values in audits or comparing several sources, it additionally should be defined what is meant by agreement. Espetvedt et al. (2013) started the comparison by defining what full agreement, minor disagreement, major disagreement, and the presence only in another system meant and how it was calculated. Box et al. (2013) calculated the percentage of key attributes that were included in the laboratory report for a random sample of data. Finally, the rate of records recorded in both systems is calculated. Bray and Parkin (2009b) state the percentage of cases that were not recorded on the database in question should be calculated.

For the capture-recapture method, the situation between two sources is presented in table 7. There are four groups of records: those that can be found from both sources (n_{11}), those that can be found only from the first (n_{10}) or the second (n_{01}) source, and finally those that are missing from both sources (n_{00}). (Bray et al. 2009b)

Table 7. Registration of records in two sources (Bray et al. 2009b)

		Source 2	
		Yes	**No**
Source 1	**Yes**	n_{11}	n_{10}
	No	n_{01}	n_{00}

When the number of articles in the three groups are identified, the estimate of records missing from both sources can be estimated. Bray et al. (2009b) presents the formula to estimate the records that are missing from both sources as

$$\hat{n}_{00} = \frac{n_{10} \cdot n_{01}}{n_{11}} \tag{10}$$

and the estimate on the total number of records is given by

$$\hat{n}_{++} = (n_{11} + n_{10}) \cdot \frac{(n_{01} + n_{11})}{n_{11}} \tag{11}$$

and thus, the completeness estimate is given by

$$comp_{estimated} = \frac{n_{11} + n_{10} + n_{01}}{\hat{n}_{++}} \tag{12}$$

Bray et al. (2009b) also presented the DC and M:I method. The DC & M:I method could be used to estimate the unregistered cancers at patients that are still alive d by assuming the proportion of unregistered cancers that caused death is the same as the proportion of registered cancers that caused death. The unregistered cases that caused death are the records that are only identified via death certificate without no mention of cancer before that. Thus, the amount of missing cases is given by

$$d = \frac{b \cdot c}{a} \tag{13}$$

where a is the number of cases registered during life which finally caused death, b is the number of cases registered during life which did not cause death, and c is the number of cases which were not registered during life but only traced via death certificate. The completeness is thus given by

$$comp_{estimated} = \frac{a+b+c}{a+b+c+d} \tag{14}$$

To estimate the completeness with this method, the proportion of cases registered during life which finally caused death is needed. The proportion should be registered independently of a death certificate. M:I ratio provides an approximation of this quantity. Even if the M:I ratio includes death certificate cases, the amount of them is usually relatively small (<10%) thus the ratio can be used as an estimate. (Bray et al. 2009b) Ajiki, Oshima, Tsukuma (1998) then present a formula to estimate the completeness in the DC and M:I method which is given by

$$registation\ rate = \frac{(1 - DC\% \cdot \frac{1}{M{:}I\ ratio})}{(1 - DC\%)} \tag{15}$$

where *DC%* refers the percentage of cases recorded by death certificate and *M:I ratio* to the mortality:incidence ratio. Finally, Bray et al. (2009b) presented how completeness could be calculated based on the Flow method. In comparison to the DC and M:I method, the Flow method assumes that in addition of a, b and c there are two groups of data missing:

unregistered cases that did not cause death (missing cases *M*) and unregistered cases that caused death but cancer was not mentioned on the death certificate (lost cases L). The rate of missing cases is then given by

$$M = s(t_i) \cdot u(t_i) \quad (16)$$

where $s(t_i)$ is the probability of surviving different intervals after diagnosis and $u(t_i)$ is the probability that patient that didn't survive had not been registered at different intervals post-diagnosis. The percentage of lost cases is given by

$$L = [s(t_i) - s(t_{i+1})] \cdot [1 - m(t_i)] \cdot [u(t_i)] \quad (17)$$

where $1\text{-}m(t_i)$ is the probability that the death certificate did not mention cancer at different intervals post-diagnosis. The completeness at time *T* could be then calculated as

$$C(T) = 1 - M(T) - L(T) \quad (18)$$

where *M(T)* presents the missing cases at time *T* and *L(T)* the lost cases at time *T*. (Bray et al. 2009b)

Some methods such as value rule analysis and source analysis are not measurable with a simple formula. When completeness is assessed using those methods, field specialists should analyze the results and their correctness. The results only give indication on the completeness and cannot be used to present the level of completeness without further analysis.

4.3 Accuracy of data

Funk et al. (2006) state accuracy refers to whether data is correct. Thus, for calculating the accuracy of a database, a definition for an error is needed (Funk et al. 2006). Baesens et al. (2010) refers to accuracy as whether a data value a' gives a correct representation of the real-world object a. They give an example of an inaccuracy in the banking field as failing to recognize the difference between AA- and AAneg ratings. Even if they are both valid values, they have different meanings. Inaccuracies occur if the incorrect representation is used. (Baesens et al. 2010) Olson (2013) notes accuracy depends on the use case of data. The level of accuracy requirements would differ a lot depending if the data was used for the development of a new surgical device or in a marketing campaign. For 85% accurate database the former would have poor data quality since human lives are concerned. In contrary, the latter would have high accuracy since any successful campaign outcomes would be considered as success even if there would be unsuccessful outcomes as well. (Olson 2013) There existed no generally accepted practice in the literature on what should be the specific level of high-accuracy data.

Different researchers see data accuracy as an important part of data quality and named other dimensions under the term of accuracy. Olson (2003) lists data accuracy as the most visible and important part of data quality. He talks about completeness, consistency, timeliness and validity under the term of accuracy. (Olson 2003) Batini and Scannapieca (2016a) name currency, volatility, and timeless being part of accuracy. It could thus be stated inaccuracies are generated in many different ways.

Olson (2003) points out six main reasons for data inaccuracies to occur: entering wrong values, carelessness, confusing data entry screens or forms, procedures that allow missing values, policies that encourage entering wrong values, or poorly defined systems. Quality improvements can be found when the root causes are identified. Olson emphasizes that a value is also inaccurate if the end user cannot interpret it. In addition, the form of the attribute can lead to inaccuracies. For example, it could be difficult to judge whether a date was invalid or only represented in another format. (Olson 2003) As there are many ways for a value to be inaccurate measuring accuracy is not always straightforward.

4.3.1 Assessing techniques

Olson (2003) names structure analysis as a method to identify inaccuracies. The analysis should be conducted by database professionals that understand the basic concepts of relational databases. Structure analysis is used to scan for values that do not obey the rules on how attributess relate to others, and how tables relate to other tables. Structure analysis deals with primary keys, primary/foreign key pairs, redundant data attributes, attribute synonyms, and other referential constraints. In addition, it deals with denormalized data sources. Structure analysis is especially important when data is moved to somewhere else, mapped in to other structure or merged since inaccuricies occur easily in these situations. For example in denormalized tables, same data is repeated which raises the likelihood of inaccuracies. When there are duplicate attributes, some of the data might be updated and some not. Information on the structures could be acquired from organizations' documentation or manuals, metadata repositories, data models or database definitions or derived by common-sense assessment. (Olson 2003)

Structures include finding functional dependencies, synonym attributes across tables and classifying relationships between tables. Structure violations can occur from violations of the structure of primary keys, denormalized keys, derived attributes, or primary key/foreign key pairs. Primary key violations mean that there are two or more records that have the same value for the primary key. For other dependencies there can be several inaccurate values. For derived attributes, a specific formula determines the accuracy. Attributes that are synonyms to each other offer different opportunities to violate the rules. In the case there are data from multiple sources, the structure analysis identifies whether data can be correctly aggregated for the target database. When the structures are identified, data values should be mined in order to find differences between the documented structures that should exists and the real structures. That way, the analyst can find new tests or the documented incorrect structure can be corrected. Finally, the defined structure is used to find inaccuracies. All the structure violations indicate inaccurate data, but the exact data records of them cannot be identified straight-forwardly since structure analysis deals with multiple values in an attribute or values in multiple attributes. (Olson 2003) Baesens et al. (2010) state that especially data flow processes are important for data quality. The data flow processes from source to end should be identified (Baesens et al. 2010).

Borek et al. (2013) stated the accuracy could be measured by defining a “gold standard” to which the values were compared. In order to measure accuracy by this technique, the reference value is needed as input. The reference value is supposed to represent the ‘real’ value. (Borek et al. 2013) Sadiq et al. (2014) compared data against manufacturer master data and the value was considered accurate if it didn’t vastly differ.

Many studies included reabstracting and recoding audit methods in which the accuracy was measured as the agreement with the original source information such as the medical records and the database values. The audit was done to a randomly selected sample of records and/or during a specific time interval. Lim et al. (2018) calculated the accuracy by transferring data to an electronic version and calculating the percentage of the differences in the double-entered data. Habibi et al. (2016) checked whether the values recorded digitally and on paper matched. They additionally named accuracy errors as records that were hard to read, incomplete or errors in sources such as bills, information that has changed lately but have not been updated, value being present when it was not supposed to be. They didn’t present any metrics for calculating the accuracy but were mostly concerned with manual checking for only the elements that users saw as problematic. (Habibi et al. 2016) In the banking context, the amount of data is so large it would be impossible to do visual inspection in all of these cases. Chen et al. (2015) also used the audit method with predefined audit rules but measured the accuracy only as how well the accuracy increased after the manual repair of errors in the original source data.

Olson (2003) names data rule analysis as a method to find inaccuracies. Data rules define generally the conditions that must always hold true for a single attribute, or multiple attributes in a table or between tables. The data rules assess data in a static state. Simple data rule analysis is used to examine whether values across several attributes can be accepted as combinations of values. These are defined as data rules. Data rules specify the constraints that must hold true across one or several attributes. Business rules are converted to executable logic that is used to test the values. Since more than one value is included in the analysis, it is not possible to say which of the values is the incorrect one. Complex data rule analysis is also used to examine values across several attributes but the data rules are more complex. These data rules check contraints over multiple business objects. Therefore, the amount of data needed for testing is greater. (Olson 2003)

Data rules can be divided to soft rules and hard rules. Data values that violate hard rules are always considered inaccuracies. If a data value violates a soft rule, it is highly likely to be inaccurate but in some cases it could be accurate. It is still important to consider soft rules because otherwise many inaccuracies could be disregarded. The rule-setting is a trade-off between setting tight rules and missing exceptions, and setting loose rules and missing inaccuracies. Data rules can be found from application source code, database stored procedures, application business procedures, or by assessing appropriate rules with a group of field experts. Many rules exist but are not controlled by the application since they can be only expressed as instructions to data entry personnels. Rules could include for example the ordering of dates, duration between events happening, or deriving a value through a business policy. When new rules are made with a group of specialists, the rules gathered from other sources should be reviewed as well. When all the rules are defined, they should be validated by looking at the rule violations. It might be that the rules were wrongly formulated or the analysis outputs indicating inaccurate data uncover new information. It is likely that there is a large number of possible rules thus it is important to choose the most important ones. Testing of all of the rules can be very time consuming and costly. When data is combined from more than one system, it is also important to consider how the rules are applicable. (Olson 2003) In the banking context, Liu et al. (2014) argued different business criterion such as the vacant ratio, invalid ratio, error ratio could be calculated and they should be in a given range. Borek et al. (2010) presented lexical analysis which was previously discussed in the overall data quality assessment methods. Lexical analysis algorithms could be used for example to find spelling errors and to analyze the correctness of text formatting from string attributes. Lexical analysis matches unstructured content to a set of structured attributes. (Borek et al. 2010)

Ezell et al. (2014) assessed the accuracy for some record elements by analyzing the relation of the value to the values of other elements. Technicians evaluated whether a value could be accurate based on the values of other elements. For example, if the serial number of a subcomponent was recorded as the same as the serial number of the overall engine the serial number attribute value was considered inaccurate. (Ezell et al. 2014)

Table 8. Summary of the reviewed methods to assess the accuracy of data

Method	Description
Structure analysis	Analyzing the violations of rules based on relational database structures
Comparing to “gold standard”	Defining a reference value that represents the ‘real’ value and assessing whether the data matches the ‘real’ values
Recoding audit	Recoding original information and comparing the database to the recoded values
Data rule analysis	Defining soft and hard data rules and assessing the violations of the rules defined
Lexical analysis	Matching unstructured string values to a set of structured attributes

The techniques for accuracy assessment discussed this thesis were divided into five method categories. A summary of the methods is presented in table 8. The table describe shortly the idea of each of the methods.

4.3.2 Measuring techniques

Ezell et al. (2014) created a binary variable to represent whether an attribute value was accurate or not. They measured inaccuracy of 7 attributes from the total of 14 attributes. The inaccuracy was defined as

$$IA_{ij} = \begin{cases} 0 & if\ the\ value\ is\ accurate \\ 1 & if\ the\ value\ is\ inaccurate \end{cases} \tag{19}$$

for *i=1, ..., 7* attributes within *j=1, ..., N_R* part records. They then estimated the proportion of inaccurate values from a sample data. For one attribute, no inaccuracies could be found from the sample. When inaccuracies were not found, a Bayes estimator (formula 8) was used. (Ezell et al. 2014) In the case of inaccuracy, *N* in the previously presented formula (8) presents the number of inaccurate values instead of incomplete values, and *a* and *b* presents the prior numbers of values that are inaccurate and accurate. If inaccuracies were found, they calculated the maximum likelihood estimates (formula 9). (Ezell et al. 2014) Habibi et al. (2016) measured accuracy by the simple ratio which was given by

$$Accuracy = 1 - \frac{Number\ of\ inaccurate\ records}{Total\ number\ of\ records} \tag{20}$$

Funk et al. (2006) presents a similar metric for accuracy. When reabstracting and recoding audit methods were used as accuracy measures, the acceptable level of agreement needed to be defined. Arts et al. (2002) state that in the case of categorical data, a value is inaccurate if it is not exactly the same as the "gold standard" value. For numerical data, an acceptable level of deviation should be defined. For example, they defined systolic blood pressure value as inaccurate if it differed more than 10mmHg from its "real" (the "gold standard") value. (Arts et al. 2002) The agreement was in many studies defined by using statistical methods with a predefined confidence limit. (Barker et al. 2012; Blevins et al. 2012; Clayton et al. 2013 ; Lim et al. 2018) Barker et al. (2012) and Blevins et al. (2012) considered a value to be accurate if it was within +/- 10% the value recorded. Before calculating the results of the measure, Barker et al. (2012) excluded the values that had a greater difference than 1000% to their comparator. They considered those values rather errors than inaccuracies (Barker et al. 2012).

Li et al. (2014) presented an approach to measure the accuracy of a relational database. They argue different datatypes are rarely considered in the current literature. First, they classified attributes into three categories: measurable attributes, comparable attributes, and category attributes. They propose an accuracy metric average relative error (ARE). For absolute accuracy, in other words the accuracy of the whole data set, the mean of error rate of different data types of attributes is given as

$$ARE = \frac{\sum_{i=1}^{card(T)}(1 - accuracy_i)}{card(T)} \tag{21}$$

where T is the set of attributes, and $accuracy_i$ presents the accuracy of attribute i. Even when the accuracy of the whole data is low, the data obtained as query results may be highly accurate. They define query result’s accuracy as relative accuracy. The authors introduce an analysis of the basic query operations and calculate the precision, recall and F-measure of a query. This thesis includes only the absolute accuracy measures, the accuracy of queries is out of the scope of this thesis. For estimating accuracy, they presented two methods: estimating the average error with and without the true values. For the case when the true value is known, the relative error of a value θ is presented as

$$RE(\theta) = \frac{\left|\hat{\theta} - \theta\right|}{|\theta|} \tag{22}$$

where $\hat{\theta}$ denotes the estimate of the value θ. The ARE of an attribute is given as

$$ARE(D_i) = 1 - \frac{\sum_{v}^{card(D_i)} RE(v)}{card(D_i)} \tag{23}$$

where D_i denotes the set of the attribute values for attribute *i*, *v* is the value belonging to D_i. The evaluation method is presented for each of the different data types where they discuss how the difference of estimates and true values is computed for each data attribute category. In many cases, the true value is not known thus the accuracy needs to be estimated with the existing values. The accuracy computation methods are different for each data type. (Li et al. 2014) The estimation algorithms are not presented in this thesis.

Fisher et al. (2009) argued a simple percentage representing the proportion of inaccurate values is not enough since the error rate does not give any indication of the complexity of the quality problems. Inaccuracies can be either systematic errors or random errors. The same error rate might not be valued the same. A problem of systematic errors where data could be wrong in only one attribute during a specific period of time may be simple to fix. In contrary, a database that has the same percentage of errors might have the errors distributed randomly across many attibutes and records which would be a lot more difficult to fix. Thus, they argue an error rate is not enough to indicate the accuracy of a database and proposed an extented measure. Their accuracy metric included calculating the error rate, error randomness measure and error probability distribution statistics. The Lempel and Ziv complexity measure was named as a possible randomness measure, and the Poisson distribution as a possible probability distribution measure. Finally, the total accuracy is given as $Accuracy = \{accuracy\ rate, randomness, statistic\}$. The probability statistics should give indication of the probability of error in any given record, the probability of error less that a decided level for any given record, and the probability of most likely number of errors in any given record. (Fisher et al. 2009)

Anderka et al. (2015) assessed the accuracy for four indications of: verification procedures, verification scope, level of expertise of the verifier, and process of quality checks. They assigned each indicator a level of quality based on predefined rules (Anderka et al. 2015). Holden (1996) assigned ratings to different categories: 0 representing poor or insufficiently registered data and 3 representing optimal data. The results were then combined to get the overall quality index. The overall quality index is the average of all ratings for each component. (Holden 1996) When the accuracy of one attribute was calculated, different methods were used to assess the overall accuracy of the data. Blevins et al. (2012) calculated the opportunity-based composite measure to combine the accuracy rate results from all data values into a single value. The composite measure was calculated by dividing the total number of accurate data values by the total number of possible data values.

Not all the accuracy assessment methods can be quantified without further analysis. If accuracy is assessed by using structural analysis, some additional analysis is needed to find the exact records with inaccurate values. In the case that data rule analysis consists of soft rules, the cases of rule violations need to be checked. When the inaccurate values are located, the number of inaccurate values and the accuracy rate can be calculated.

4.4 Consistency of data

Inconsistencies can occur between different sources and between values in one source. The values, rules and formats should be consistent across different data sources. Consistency was understood in several ways in the researched articles. Amoroso et al. (2014) used the term internal consistency and external consistency as stated in the World Health Organization (WHO) framework. Baesens et al. (2010) discussed interrelational consistency which referred to consistency between all the records in a dataset. They also discussed intrarelational consistency which referred to consistency of one record. Additionally, they referred to external consistency while discussing the importance of values being consistent within branches. In the credit modelling context, it is very important for the default information to be consistent for all branches. If a client is in default in one branch but not in the rest, the risk for future losses increases. (Baesens et al. 2010) In this thesis, all were included: internal consistency between values of a record and between records of a dataset, and external consistency between datasets.

Data edits are an important part of consistency, according to Batini and Scannapieca (2016a). They define data editing as the task of revealing inconsistencies by formulating rules that must be respected by each true set of answers. They understand consistency, cohesion, and coherence being the capability of the values to comply without contradictions to every real-world event or object, as specified in terms of integrity constraints, data edits, business rules, and other formalisms. Olson (2003) advocates that inconsistencies create inaccuracies even if the values are correct and the user could interpret them as same values. Inconsistent values, for example a city written in two different ways for different records, cannot be accurately aggregated and compared. If the data is later used for a new unintended use, inconsistencies could create inaccuracies.

4.4.1 Assessing techniques

Amoroso et al. (2014) assessed the data quality of national health management information systems. Consistency was assessed in three different ways. First, they searched for extreme and moderate outliers across all 10 indicators chosen. The indicators were chosen based on WHO recommendations and priorities. The data was aggregated at facility-level for each month. For a specific region, the average value of the indicator was calculated for a specific time period. The monthly values that differed at least two standard deviations from the average were considered as moderate outliers and those that differed at least three standard deviations were considered as extreme outliers. Consistency was then defined as the absence of extreme outliers. (Amoroso et al. 2014)

Amoroso et al. (2014) additionally measured internal consistency by assessing two different ratios recommended in WHO framework. The ratios were calculated over time and both assessed one indicator compared to another. The level of inconsistency was then set to a specific difference percentage in the ratios. For the first ratio calculated, consistency was defined as the district ratio differing less than 33% from the national ratio. In the second ratio, consistency was defined as a specific indicator being less than 2% greater than the other indicator. Finally, they assessed consistency over time by calculating the ratio of number of events during a specific year and the mean number of events during previous three years. Again, consistency was defined as the district ratio differing less than 33% from the national ratio. (Amoroso et al. 2014)

Ezell et al. (2014) examined an aircraft maintenance data base. 13 data properties were used to calculate inconsistency of each record. They defined business rules for attributes to measure whether the values comply with them. The business rules included rules for one element or rules for the relationship of several elements. For example, a serial number was considered inconsistent if it was 8 or 10 characters long. (Ezell et al. 2014) Also Habibi et al. (2016) examined consistency by searching for syntax violation to assure for example unit consistency. They also searched for duplicate data (Habibi et al. 2016).

Gray et al. (2015) assessed data quality of a longitudinal study of adolescent health. They assessed the internal consistency by observing the expected relations within a data set that measured similar traits. For example, they assessed the relation of allergy status and the prevalence of asthma in the data collected. They then analyzed whether the observed associations were reasonable. (Gray et al. 2015)

External consistency could be assessed by comparing values between sources. Liu et al. (2014) talked about measuring the consistency of business indices such as vacant ratio, invalid ratio and error rate between systems. Sadiq et al. (2014) implemented a prototype based on their query answering data quality framework. As a consistency measure, they examined whether a value is the same in another system. For a manufacturer attribute, the prototype assessed whether the manufacturer name matched a master data source with all the correct manufacturer names. (Sadiq et al. 2014)

Table 9. Summary of the reviewed methods to assess the consistency of data

Method	Description
Historical ratios	Analyzing the reasonability of ratios
Outlier detection	Calculating the average and the standard deviation of values and identifying the extreme values that deviate from others
Syntax violation	Identifying violations of data rules
Expected relations	Analyzing whether the observed relations of values are reasonable
Comparing several systems	Linking records between two or several databases and identifying the values that do not match

The techniques that were used to assess completeness in the literature discussed in this thesis were summarized in five method categories. A summary of the methods and a short description of them is presented in table 9.

4.4.2 Measuring techniques

Ezell et al. (2014) created a binary variable to present whether an attribute value was inconsistent or not. They measured inconsistency of 13 attributes from the total of 14 attributes included in their study. The inconsistency was defined as

$$ICN_{ij} = \begin{cases} 0 & if\ the\ value\ is\ consistent \\ 1 & if\ the\ value\ is\ inconsistent \end{cases} \tag{24}$$

for $i=1,\ldots,13$ attributes within $j=1,\ldots,N_R$ part records. They then estimated the proportion of inconsistent values from a sample data. For 7 attributes, no inconsistencies could be found from the sample. When inconsistencies were not found, a Bayes estimator (formula 8) was used. In the case of inconsistency, N in the previously presented formula (8) presents the number of inconsistent values instead of incomplete values, and a and b presents the prior numbers of values that are inconsistent and consistent. If inconsistencies were found, they calculated the maximum likelihood estimates (formula 9). (Ezell et al. 2014) Amoroso et al. (2014) calculated the proportion of values that were considered inconsistent based on their predefined rules. The proportion was calculated as the number of inconsistency occurrences divided by the total. For outliers, the results of each indicator were combined as one quality percentage. The percentage of inconsistencies was calculated as the sum of occurrences for all the indicators divided by the total number of values in all the indicators. (Amoroso et al. 2014)

Gray et al. (2015) assessed the internal consistency by observing the expected relations within a data set that measured similar traits. They used bivariable log-binomial models, Poisson-distributed generalized models and bivariable linear models to examine the relationships between attribute values. (Gray et al. 2015) The formula was not presented in their study.

4.5 Timeliness of data

Liu et al. (2014) stated the timeliness included the time taken for transmission and extract-transform-load procedure (ETL) processing. In the credit risk modelling context, Baesens et al. (2010) discusses timeliness presents the data being recent in relation to its intended use. Having recent data is important but it is equally important to having it at the exact time they are needed (Baesens et al. 2010).

Olson (2013) remarks timeliness requirements depend on the intended use of data. If sales data is updated after the end of each month, it is poor-quality data for computing sales bonuses that are due to the end of the month but high-quality for historical trend analysis if the end user knows when the system is updated. The same data might be poor-quality for short-term use but high-quality for long-term decision making. (Olson 2013) Bray and Parkin (2009a) point out there is a trade-off between the timeliness and the other measures of quality. Databases might also have predefined time intervals for updating data. (Bray and Parkin 2009a)

Sadiq et al. (2014) gave an example that the update of prices should be seen within one month from the data. In the medical field, timeliness was mostly measured as the time between an incident, procedure or diagnosis, and the registration or reporting (Asterkvist et al. 2019; Box et al. 2012; Bray et al. 2009; Bray et al. 2018; Busacker et al. 2017; Heikkinen et al. 2017; Lim et al. 2018; Jonasson et al. 2012). The registration or reporting date indicated the availability to users. The timeliness was then assessed over a certain period of time. Asterkvist et al. (2019) examined the timeliness by calculating the difference between the first day of diagnosis and the reporting date. They conducted an audit of three months. They created a graph representing the percentage of cases registered over months since diagnosis and assessed the timeliness during a given number of months since diagnosis. (Asterkvist et al. 2019) Lambe et al. (2015) assessed timeliness by comparing the registration status and date of registration to a database in which registration was mandatory by law.

As the definition of data quality suggested, the level of high-quality depends on the use-case. The results of this study support that as high-quality timeliness differed vastly in the reviewed studies. Busacker et al. (2017) defined data as timely if the event was recorded within 10 days. Lim et al. (2018) considered data as timely if data was available in 12 months or less. Jonasson et al. (2012) calculated three timeliness levels: the proportion of cases registered within one year, the proportion of cases registered within 15 months, and the proportion of cases registered with two years. In the researched literature, no consensus on the optimal level of timeliness existed.

Anderka et al. (2014) measured timeliness using two metrics by making a distinction between a “core” list and a “recommended” list. They defined three levels of quality where the optimal was greater or equal of 99% meaning 99% or more of the records were completed within 2 years, the second level greater or equal to 95% and the last level greater or equal to 75%. (Anderka et al. 2014) Box et al. (2013) presented timeliness using the Kaplan-Meier curves which present the percentage of unregistered notes over the amount of time passed.

European Central Bank (2018a) defined timeliness as not only up-to-date but also current thus currency measures are included. Heinrich and Klier (2011) researched the measuring methods for currency of data but it was included as timeliness measure also since they used the term timeliness as well. They argued currency is part of timeliness. They presented different formulas and named six requirements for currency measure presented in the literature: normalization, interval scale, interpretability, aggregation, adaptivity, and feasibility. They then proposed a probability-based metric for data currency which is given as

$$Q_{curr}(v_i) := \exp\left(-decline(i) \cdot age(v_i)\right) \tag{25}$$

where v is an attribute value of an attribute i, *decline(i)* is the average decline rate of the shelf life of the attribute values of the attribute i, and *age*(v_i) is the age of the attribute value. If the value is acquired instantly when assessing data quality, *age*(v_i)*=0* and then Q_{curr}=1 meaning the value is up-to-date. The metric could be applied to a single record by considering the weighted arithmetical mean of the values of the metric of attribute values. This is defined as

$$Q_{curr}(\tau_j) := \frac{\sum_{i=1}^{card(T)} Q_{curr}(\tau_{ij}) g_i}{\sum_{i=1}^{card(T)} g_i}, \quad where\ g_i \epsilon [0; 1] \tag{26}$$

where τ_j is a record with a certain number card(T) of attributes, τ_{ij} is the value of the *i*th value in the *j*th record and g_i represents the weighting of attributes. They finally applied the metric in mobile services campaign management. (Heinrich and Klier 2011)

The timeliness dimension and its assessment were mostly agreed in the literature. The only method to assess timeliness was to calculate the time taken for the data to be recorded in the database or to be accessible to data users. The results could be presented as timeliness level of the database where the percentage of timely data indicates the quality level, as a single percentage score of data recorded after a specific time period or as a graph that visualized the percentage of registration over the amount of time passed.

4.6 Uniqueness of data

The uniqueness was not mentioned as a measured dimension except in one paper. In the credit risk modelling context, uniqueness is a very important dimension. Baesens et al. (2010) give an example of problems caused by duplicate records. If there exists duplicates in loan agreement data, the risk of credit loss is calculated twice and thus twice the regulatory capital is needed. Such errors could consume unnecessary capital. (Baesens et al. 2010) Liu et al. (2014) discussed about uniqueness. They established a data quality evaluation system for a commercial bank. They measured uniqueness as not finding duplicate primary keys in a table. (Liu et al. 2014)

There were some papers that discussed the identification of duplicate records without mentioning the uniqueness dimension (Liaw et al. 2014; Bray and Parking 2009b; Olson 2003). Those measuring methods are also included in this subchapter. Liaw et al. (2015) assessed the duplication of patient records from data extracted for one year every four months. They calculated the number of duplicate records and the percentage of duplicate records. Olson (2003) discussed analyzing and understanding primary key constraints and properties can give indication on the confidence of data being unique.

Some of the techniques that were used to measure completeness were named also as techniques to measure uniqueness. Baesens et al. (2010) state uniqueness is an additional measure to completeness. As completeness examines the under-registration of data, uniqueness examines the over-registration of data (Bray and Parkin 2009b). Thus, the similar techniques that give indication on completeness could also give indication on uniqueness. These techniques include the historical data methods (Bray and Parkin 2009b) or the value rule analysis when for example values have unreasonably high frequency (Olson 2003). Bray and Parkin (2009b) discussed that assessing historical trends be used as an indication of duplicate records. Systematic inconsistencies (over multiple sites) in incident rates can be an indication for over-registration. (Bray and Parkin 2009b)

Additionally, data quality assessment techniques that were presented as methods to assess the overall quality of data discussed value uniqueness or duplicate values. One method was presented by Borek et al. (2011). They listed matching algorithms such as probabilistic matching techniques, as a method to identify duplicate records. The algorithms check multiple attributes and records to detect matches. (Borek et al. 2011) The matching algorithms were not presented in their paper.

Table 10. Summary of the reviewed methods to assess the uniqueness of data

Method	Description
Primary key attribute properties	Examining the primary key properties and analyzing the violations of the constraints
Value rule analysis	Searching for unreasonable results through data aggregation, assessing historical trends such as incidence rates
Matching algorithms	Detecting value matches from attributes and records to identify duplicate values

The uniqueness dimension has not been widely discussed in the literature. Still, the articles discussed values being unique or values having duplicates in the dataset. A summary of the methods discussed in this subchapter are presented in table 10. The table describes shortly the techniques belonging to each of the methods.

4.7 Validity of data

Bray and Parkin (2009a) used the term validity but also concluded it is a synonym for accuracy. Olson (2003) used the term validity under the term of accuracy. Validity of value refers to the value having correct form independent of the real value while as accuracy refers to values having the value as close as possible to the real value. For example, for a color attribute a value that indicates blue would be correct, but it can still be inaccurate if the real value would be green. (Olson 2003)

Olson (2003) discusses that the validity of values can be examined by analyzing column properties. Column properties are also referred as domain definitions, and they can be seen as value rules. Column property analysis examines single attributes independent of all other attributes. The values can be investigated by comparing them to specific constraints to that specific attribute. The more constraints included, the larger the probability to identify the possible invalid values. (Olson 2003)

First, the column properties should be defined. The properties tell what values are considered acceptable. Information on the properties could be gathered from the database and data entry screen specifications, data entry procedure and data dictionary documents or manuals, and metadata repository information. The properties include business meaning, storage properties, valid value properties, empty condition rules and other descriptive information. Business meaning of an attribute tells what should be stored in it. Yet, it is not always the case since in practice an attribute can be used for another purpose or not be used at all. Storage properties include rules about data type, length, and precision for numeric attributes. These basic rules are usually forced by the database structure but violations can still be found. For example, noncharacter data saved as character, 30-character attribute always getting two-character values, name attribute getting 1-character lenghts, or only integer values in data type decimal or float. Examples of typical column properties and examples of invalid values are given in table 11. Valid value properties specify the acceptable values. The properties can include a discrete value list, range of values, skip-over rules, text-attribute rules, character patterns or special domains. Empty condition rules examine whether the NULL values are allowed. If there shouldn't be any NULL values, it should be analyzed whether the values contain things such as question marks, blanks, or values such as "none", "not known", "not applicable". Other descriptive information could be any information that helps to identify the probability of values being invalid, such as if an attribute is forced to be unique and not null, the likelihood of previously mentioned codified NULL values is small. Another example could be a database system that only accepts valid date values. The properties might seem as self-evident but analyzing them could for example reveal changes over time, transformation problems between sources or transformation problems caused by combining data from multiple sources. (Olson 2003) The properties that could be included in the analysis and possible invalidities are presented in table 11.

When the column properties are defined, data should be discovered independently of the properties in order to avoid bias. Data properties should be discovered and then compared to the documented properties. It then allows the user to see where there's an error in the data or where the documented properties are either invalid or incomplete. Finally when the documented properties are checked, the violations of the defined rules are searched. All the values that violate the rules are invalid values. It is important to notice, the column property analysis does not find values that are valid but incorrect, nor the invalid values you don't have a rule for. (Olson 2003) Finally, values are considered as valid if they do not violate the rules.

Table 11. Examples of column properties based on Olson (2003)

Property	Examples of possible invalid values
Business meaning	The attribute not containing the information it should contain
Data type (e.g. Character, attribute character, integer, small integer, decimal, float, double precision, date, time, timestamp, binary, doublebyte)	Storing noncharacter data in character type attribute has allowed users to enter same information in multiple formats
Numeric precision	The precision of numerical value is not correct
Character set, code page	System code differences generating invalid values when moving to another system
Length restrictions (shortest, longest, variability) Distribution of lengths	If an attribute that should contain names has 1-character values
Acceptable values - discrete list of acceptable values (encoded value meaning) - range of acceptable values (and skip-over rules) - text attribute restrictions - character patterns - character exclusions Distribution of values	Text attribute indicating city includes special characters Phone number following a pattern of 9999999999 Entry person entering own birthday when the birthday of a customer was not known, high frequency of one date
Null rule	Blanks, question marks, special characters, different texts such as “none”, “not known”, “not applicable”
Unique rule	Duplicate values
Consecutive rule	Missing values between the lowest and highest values

Bray and Parkin (2009a) researched the different methods to assess the validity of data in the medical field. Validity could be measured by linking records in several databases and comparing the values to see if they match. Comparison could also be done by doing comparison of the values within a database, within a subset of data or comparing values over time. (Bray and Parkin 2009a) In the literature, the validity was assessed by comparing two or several sources (Box et al. 2013; Lambe et al. 2017). Box et al. (2013) calculated validity as the agreement between the elements in the two systems. They collected attribute values from one system and their comparators from the another system, and then compared the values, and calculated the significance of agreement using statistical tests. (Box et al. 2013) Lambe et al. (2017) also compared data values across multiple source. Gray et al. (2015) talked about external validity. They assessed the validity of predictive variables of lung function by comparing their result to the values expected based on the literature in the field. (Gray et al. 2015)

Bray and Parkin (2009a) named reabstracting and recoding audits as the most objective method to assess validity. The audit could be used to assess the differences between paper records and database values, or differences between the work of different data collectors. The objective in reabstraction is to reabstract and code data from the source by experts, and then calculate the extent of agreement between the source data and coded data. Recoding is a similar process but the source documents are not reviewed in the process. The reliability could be assessed by testing the understanding of coding rules. (Bray and Parkin 2009a) The audits were used in several papers. They then used statistical tests to calculate the level of agreement. (Arboe et al. 2016; Asterkvist et al 2019; Bah et al. 2013; Lambe et al. 2015). Arboe et al. (2016) assessed the validity by crosschecking the medical records to the database for subgroups of patients. Asterkvist et al. (2019) assessed validity having field experts comparing the data values to medical records for a sample of patients.

Bray and Parkin (2009a) also argue the number of unknown/missing values in a record could give indication of the data quality validation. They state unknown values can be caused by system problems, source document access problems, value definition problems or misapplication of coding rules. When some important values in a record are missing, it is more probable that the other values are not valid. (Bray and Parkin 2009a) Bah et al. (2013) argued the cases with unspecified sites and with unknown age affected the validity of data. Bray et al. (2009), Bray et al. (2018), Heikkinen et al. (2017) and Jonasson et al. (2012) assessed the number of cancer cases with unknown primary site by age group and site and compared them with the chosen registries with chosen statistical tests. Unknown primary site means it is not known from which body part the cancer first started.

Bray and Parkin (2009a) also argue internal consistency can be used as a method to understand the data validity. They argue validity could be assessed with logical rules for single attributes, several attributes in a single record, several attributes in several records, or several attributes within several databases. Rare exceptions can violate the predefined rules, but the value is still valid. If a case has been verified earlier to be correct, it should be assigned an override flag so that different users don’t have to validate the same value several times. (Bray and Parkin 2009a)

Olson additionally discussed value rule analysis as a possible method to find invalid values. Value rule analysis is used for trying to find unreasonable results through cardinality, counts, sums, averages, medians, frequency distributions, standard deviations and other similar aggregations. Bray and Parkin (2009a) also named historical verification as a validity assessment method. The percentage of morphologically verified (MV%) cases could be calculated for cancer registries since the accuracy of diagnosis is usually higher when the cases are histologically assessed. The values of MV% should be compared to the expected values. The percentage of death certificate only (DCO%) could also give indication of the accuracy in the case of cancer records since the information on death certificates very often lacks accuracy. (Bray and Parkin 2009a) Jonasson et al. (2012) assessed the MV%, DCO% and compared the results with other European countries. Bah et al. (2013) examined the ratios of different verification sources since some are considered more trustworthy than others. Bray et al. (2009), Bray et al. (2018) and Heikkinen et al. (2017) used historical verification methods and compared the results with chosen registries using statistical tests.

Table 12. Summary of the reviewed methods to assess the validity of data

Method	Description
Column property analysis	Examining the properties of single attributes independently of other attributes and analyzing the violations of the predefined properties
Comparing sources	Linking records between two or several databases and identifying the values that do not match
Recoding audit	Recoding original information and comparing the database to the recoded values
Unknown value analysis	Identifying the records with missing key information and analyzing their validity
Value rule analysis	Searching for unreasonable results through data aggregation, assessing historical trends
Logical rules	Searching for violations of predefined rules of single attributes or between attributes in a single record, in multiple record or multiple databases

The validity assessing techniques discussed in this thesis were divided into six method categories. A summary of the methods used for assessing validity is presented in table 12. Also for validity, a simple validity rate can be calculated by first measuring the number of invalid values. Asterkvist et al. (2019) used recoding methods and measured validity as the percentage of number of women recorded similarly in data and medical records divided by the total number of women. (Asterkvist et al. 2019) In order to measure the strength of agreement between the original and reabstracted date, different statistical methods were used depending on the data type (Asterkvist et al. 2019; Lambe et al. 2017). No additional measuring techniques were presented for validity only.

4.8 Summary and discussion of the literature review results

The assessment methods for different dimensions discussed in this chapter are presented in table 13. The metrics for different dimensions are overlapping for some dimensions. Thus, it could be more reasonable to first assess the most relevant metrics in the terms of data types and limitations and make conclusions for different dimensions based on the results.

Table 13. Summary of the reviewed assessment methods for all dimensions

Dimension	Methods
Completeness	Number of NULL values Rules for different levels of completeness Value rule analysis Case-finding audit Comparing sources Source analysis
Accuracy	Structure analysis Comparing to “gold standard” Recoding audit Data rule analysis Lexical analysis
Consistency	Historical ratios Outlier detection Syntax violation Expected relations Comparing several systems
Timeliness	Time taken for data being recorded Currency
Uniqueness	Primary key attribute properties Value rule analysis Matching algorithms
Validity	Column property analysis Comparing sources Recoding audit Unknown value analysis Value rule analysis Logical rules

Most of the data assessment methods that were listed in the overall quality metrics were similar to those that were discussed under specific data quality dimensions. Abela et al. (2014) discussed techniques similar to column property analysis and value rule analysis, and also used death-certificate methods, the percentage of morphologically verified, and incidence rates. Hinterberger et al. (2016) used techniques similar to column property analysis and data rule analysis. The metrics presented by Borek et al. (2011) also included measures from column property analysis, value rule analysis and audits. Matching algorithms and lexical analysis were presented as additional tests, and they were linked to uniqueness and accuracy dimensions (Borek et al. 2011). Majumbar et al. (2014) and Funk et al. (2006) presented methods for the whole data set that could be applied to each dimension depending on the indicators chosen. Majumbar et al. (2014) presented the MCDA method where assessing criteria should be defined. Funk et al. (2006) discussed a data quality survey could be conducted where the questions should be defined. Charrondiere et al. (2016) also proposed different checks based on different data rules. Interestingly, they additionally included the level of documentation into data quality assessment. For high-level data quality, comprehensive data documentation should also be available. The documentation should include for example the calculation methods used. (Charrondiere et al. 2016)

Completeness assessment methods finally included five principal methods. To assess the completeness internally in a system, assessing the rate of NULL values is a simple method. Is it straight-forward for the mandatory values but requires more analysis for the attributes that don't necessarily have a value. It should also be analyzed whether the system could be values that indicate missing values. Regardless of its defects, it is a simple method to implement and monitor. In contrary, case-finding audits can be a very demanding method to assess data quality especially when the regulation demands banking institutions to have continuous and regular assessment processes. Comparing sources can also be a demanding method if there are several source systems which is generally the case in large organizations. It would be less time-consuming to implement the comparison to only a sample data and then estimate the completeness for the whole data. Value rule analysis and source analysis can give indication on the completeness, but no results are immediately acquired. To implement these techniques to continuous monitoring some references are needed to be defined. This is similarly the case when using predefined rules for different completeness levels. The special methods that were presented in the medical field cannot be used for the credit risk modelling data without defining the application in different context.

The methods to assess accuracy were summarized to four methods. Recoding audits would be very demanding to implement especially continuously. Similarly, comparing the values to their "gold standard" values could be time-consuming in the case with multiple sources. It would less demanding to use these techniques to a sample data and then estimate the results to the whole data set. Structure analysis and data rule analysis can give indication of accuracy and in the of data rule analysis, hard rules can show the exact inaccuracies. However, these two methods would require predefined requirements to be useful. When the data assessment and rule setting phases are first conducted in detail, accuracy could easily be monitored based on those rules.

The consistency assessment methods had most disagreement whether to assess the consistency between values, system, or multiple systems. European Central Bank (2018a) defined it as data matching between different data systems inside the institution. Thus in the credit risk modelling context, the assessment methods should be chosen so that the consistency of data between sources is assessed. Comparing several systems could thus be a good practice but it can be time-consuming. It could be less demanding to compare values of a data sample and then estimate the rate for the whole data set. Additionally, using the historical ratios method could be applied so that the ratios are compared to the ratios obtained from the other systems. Outlier detection, syntax violation and expected relations assess consistency rather inside a system than between systems.

Assessing the timeliness dimensions included assessing the time taken for data be recorded and the currency of data. In credit risk modelling, it could be measured how long it takes for data to be available for the data users. The time taken usually depends on the systems and their regular update times. The modelers assessing timeliness should have a limit on when data should be accessible depending on their needs. It is important to include the estimate on the currency of data to have information on how up-to-date the loan agreement and customer data is.

The most essential phase for all dimensions is to define what does it means for a value to correct. For example, when measuring accuracy what does it means for a value or a record to be accurate in this special case. Multiple examples have been presented in the literature review but since the articles in this literature review were from different fields, the results cannot be straight-forwardly converted for credit risk data. Although, the data type gives useful indication on what kind of methods could be suitable.

The simplest method to assess uniqueness was to test whether primary key properties were violated. This would be a very simple method to monitor on a regular basis. The assessment methods of completeness such as value rule analysis could give indication on not only missing values but also duplicate values. Predefined limits should be set for value trends in order to monitor the trends automatically. Matching algorithm could be a good method to identify very similar values and duplicate results but might be demanding to adopt to very large data sets.

The validity assessment consisted of six methods which were very similar that the methods to assess other dimensions. As previously discussed, comparing sources and recoding audits could be demanding to implement and monitor and value rule analysis could only give indication of the validity of values. Column property analysis and logical rules are demanding to implement in the beginning but could be easily monitored after. When the data assessment and rule setting phases are first conducted in detail, invalid values could be easily monitored and found based on the predefined rules. Unknown value analysis could give indication on what records should especially be checked.

Many of the methods discussed emphasize the importance of understanding what the data should consists of. Even if this thesis presents specific examples of different possible assessment rules, they are from different fields and cannot be adopted in the credit risk modelling context. The results are expected since even the definition of high-quality data underlines the importance of understanding the intended use. Most of methods need predefined criteria on what is an acceptable value. To effectively adopt the methods, collaboration between business professionals and database professionals is needed. Different professionals are needed both before adaptation to define the data properties and rules, and during the analysis to inspect and elaborate the results. Also, comparison values are used in many methods. Business professionals in the field are again needed to analyze what values could be used as reference data or what are the acceptance limits, and database professionals to how the sources are linked and how data is transferred between them. It is clear that the assessment of data quality cannot be solely left to the IT department since industry knowledge is essential in the process.

All the dimensions could then be similarly presented as a simple ratio of correct values. Likewise, simple scores for different levels of correctness could be predefined. The score could be calculated for the whole data set or be estimated by using a randomly selected data sample. In the medical field, specific methods to measure completeness were proposed. Also, additional technique was proposed for measuring the accuracy dimension. The technique included measuring not only the error rate but also the complexity of errors. Some of the methods only give indication whether there might be problems in data. When using these methods, further analysis is needed.

After calculating the ratios or giving the scores, it is a matter of preference if the results are combined. The combined score could be obtained by calculating the average error, by choosing the minimum or maximum of all scores, or by calculating the weighted sum. Documenting the results and the availability of the results is an important part of data quality assessment.

It is not possible to reach perfect quality thus it is important to find the level of acceptable quality where the quality deficiencies left do not have a significant effect on the modelled credit ratings. Especially when the perfection of other dimensions may influence the quality of other dimensions. Additionally, some assessment methods could be infeasible to implement. Field experts should find the best quality evaluation methods for their purposes and concentrate on those rather than employing each possible method.

5 EMPIRICAL STUDY

Data used in this thesis was collected from a case company that is operating in banking business. The data set consisted of a sample of a few specifically selected attributes of credit loan agreement data which was collected from the same month of three consecutive years. During this time, no changes occurred in the coding of these attribute values. The set consisted of 11 attributes and around 1.85 million records in total. The data set was collected specifically for the purpose of the analysis of assessment methods, and it does not represent the real data set used for modelling purposes. In this thesis, some tests for completeness, accuracy, consistency, uniqueness and validity was tested. The test data set did not include all the needed information to assess timeliness. The tests were implemented using SQL and R.

Industry professionals' input was used to distinguish the techniques that could be applied to the test data set. The chosen attributes are presented in table 14. The actual business names and descriptions of the attributes are not given due to company requirements. This business meaning of attributes was analyzed from company documentation and data rules were formed based on the documentation and professionals' input. The data type, null rule and data category was analyzed based on data properties. These properties were analyzed from company documentation and relational table structure. For string attributes, the maximum length in characters was defined in the database structure. The character set was defined as Unicode and the collation as Finnish. For numeric values, the precision and the scale were defined. For date types, the datetime precision was defined. The primary key of the data set table consisted of the date and the ID (attribute 2) thus no duplicate agreements could exist for the same date. These restrictions set the basis for the analysis since these rules do not have to be tested.

First, the data was analyzed by calculating the frequency distributions of values and doing visual inspection on the first 1000 values in a random order. Even though attributes V4-V6 were set as string values, they consisted of numerical codes. Also attribute V11 was presented by a binary value even though it was set as an integer value. This was due to the table implementation. These properties would allow the entry of invalid data.

Table 14. Description of the attributes included in the empirical study

Attribute	Data type	NULL values accepted	Category
V1	Date	NO	Continuous
V2	Integer	NO	Continuous
V3	Date	YES	Continuous
V4	String	YES	Categorical
V5	String	YES	Categorical
V6	String	YES	Categorical
V7	String	YES	Categorical
V8	Numeric	NO	Continuous
V9	Numeric	NO	Continuous
V10	Numeric	NO	Continuous
V11	Integer	YES	Categorical

The completeness rate was calculated as the percentage of values that were not NULL. There existed no NULL values for V1-V10 for which the completeness rate was 100%. For attribute 11, measuring the completeness was more complicated. The rate of non-missing values for V11 was 15.9%. The business meaning of the attribute suggests there should be more NULL values than other values so the result is in line with the business meaning. The percentage should be compared to a reference value to analyze its' reasonability. In this test data set, the resulting rate cannot be compared to any reference value since the data set was not collected randomly thus the percentage does not reflect the real percentage.

For continuous numerical data, aggregations such as the sum, standard deviation, average and median was calculated. The results were analyzed by inspecting the trends over three years. None of the monthly results varied notably from the expected three-year trend. The value frequency distributions were calculated for each attribute. The trends were mostly consistent. Attribute V3 had a slight decrease that varied from the expected trend over the three-year period. This irregularity could be further investigated to see if there are any data quality issues occurring. The results show no indication of duplicate records occurring since there are no increasing irregularities. Case-finding audits were not possible to conduct for the empirical part of this thesis. Nor comparison between systems were possible since the data set did not represent the real data set.

The data rules were formed based on company documents, manuals and professionals' advice. For date attributes, no dates that occurred in the future were allowed. Additionally, a skip-rule was presented for attribute V3 so that the date value should not be on a weekend. Based on the data rules defined, the validity of data attribute V3 received a rate of 99.84%. Attribute V1 was 100% valid. For categorical attributes, the values were compared to company manual so that if the code existed in the company manual the value was valid. Also for V2 and V4-V7, the length was set to a specific number of characters. No invalid values could be found from other attributes than the V3. This result was expected since the formation of the test data set included only valid values for most attributes.

For accuracy, only the attribute V2 was tested. The ID numbers were compared to another data source so that all the values should also be found from the second data set. It was not possible to conduct audits. No additional data rules were formed as there was not enough attributes to compare the relation of values whether they are accurate as a group of values.

For consistency, outliers were searched for numerical attributes. Moderate outliers were defined as values being greater or less than 2 standard deviations from the average value and extreme outliers as greater or less than 3 standard deviation from the average value. The results are given so that the extreme outliers are also included in the moderate outliers. For attribute V8, 0.66% of values were considered as moderate outliers and 0.38% extreme outliers. For attribute V9, 0.17% of values were moderate outliers and 0.09% extreme outliers. Finally for attribute V10, 0.64% of values were moderate outliers and 0.47% extreme outliers. The reasonability of the outlier values should be further analyzed.

The results of the empirical study suggest that the methods discovered in this thesis are suitable for credit risk data but specific limits should be set to acceptable values. For this purpose, it was not possible to have the acceptance thresholds due to company requirements. More analysis of the acceptability of values should be conducted to truly find underlying errors from the data set. Additionally, further research should be conducted to find acceptable levels of quality levels and applications of methods in the banking context.

6 CONCLUSIONS

Organizations’ business models rely increasingly on information and high-quality data is a necessity for information quality. While information systems have advanced rapidly during the latest years appropriate data quality assessment tools have lacked behind. At the same time, new use cases for data are constantly developed and data is seen as an asset to gain strategic advantage. The issue of poor-quality data is a universal problem across different companies. Large organizations are now working towards improving their data management programs and addressing their data quality issues. Especially the banking industry which is facing increasing demands due to tightening regulation and customers’ growing demands. The financial crisis got regulators to increase their demands and supervision on banking institutions’ capital requirements and their calculation. Since internal ratings based models are based on banks’ internal data, also data quality became an important component of regulation. Due to regulatory demands, ensuring the quality of data has been a primary concern to most banking institutions since the regulation affects banks’ capital requirements. High-quality data is especially important to achieve high-quality ratings in which the decision-making of credit loans depends on. High-quality data is a case-dependent concept which refers to data being qualified and having an acceptable level for its specific purpose, in this case for credit risk modelling. High-quality data also presents precisely the real-world object or events it is supposed to be presenting.

The objective of this thesis was to identify suitable methods to assess and measure data quality for the credit risk modelling purposes. First, the regulatory requirements are discussed, and the needed data quality dimensions are defined. Then, data quality assessment and measuring methods are discovered by conducting a literature review. Assessing methods refer to methods that evaluate the condition of data. Measuring methods provide quantifiable methods that are repeatable and that can be used to compare the improvement or deterioration of data quality over time. According to regulatory demands, banks should employ solid and systematic data quality management practices that cover data from its entry to reporting. The practices should include the assessment of the completeness, accuracy, consistency, timeliness, uniqueness, validity, traceability, and availability/accessibility to comply with the regulation. In this thesis, the dimensions of completeness, accuracy, consistency, timeliness, uniqueness, and validity are covered.

The assessment methods and measuring techniques are presented and discussed in this thesis. The results of the literature review show that the implementation of data quality methods require the collaboration of experts from different fields. First, it is necessary to understand the use-case of data, what the data represents and what constitutes as erroneous data or what data is considered as the real values. Then, the quality of data can be measured. For some methods, additional analysis is needed. Field experts should find the best quality evaluation methods for their purposes and concentrate on those rather than employing each possible metric.

The regulation requires banks' to have consistent criteria and metrics with clearly set tolerance levels and thresholds, and continuous assessment procedures. The acceptable criteria, metrics and thresholds are not given by the regulation. As the results of this thesis show, neither are there generally accepted practices in the banking field. Thus, banks should have their own applicable practices based on their knowledge on their own internal data and models. The acceptable quality levels should be chosen so that the results of the modelling are not significantly affected by the remaining data quality issues. The assessment methods presented in this thesis can be useful only with well-defined acceptance levels and criteria. Some of the methods could be more easily implemented than others but the best methods depend on the database systems and the variables used. Additionally, it is important to note that regulation requires the assessment frameworks to be current. Thus, the assessment methods should be periodically revised and evaluated whether they still meet the expectations.

Most importantly, banks' should clearly define the criteria and metrics used, have clearly defined roles and solid documentation practices and apply the monitoring processes in a continuous and consistent manner. The results of this thesis could be used to improve the data quality assessment practices but further research should be conducted to find acceptable quality thresholds and applications of methods in the banking context. The banking institutions should reflect the reviewed methods into their own data and apply the methods in that context.

In order to use data as a strategic advantage, the issues of data quality should be managed proactively, and long-term improvement activities should be employed. An important part of regulatory compliance is to understand the root-causes of data quality issues. This thesis did not focus on how to identify the sources of the issues and how to improve data quality. When establishing data quality management frameworks and practices, it is essential to include these steps. An important part is to include the management of IT systems, and all the stakeholders such as executive, data entry people, and data users.

REFERENCES

Abela, L. Allemani, C. Coleman, M. P. Li, R. Moore, J. Nur, U. Rachet, B. Woods, L. M. 2014. Control of data quality for population-based cancer survival analysis. *Cancer Epidemiology*, 38(3), pp. 314–320.

Ajiki, W. Oshima, A. Tsukuma, H. 1998. Index for evaluating completeness of registration in population-based cancer registries and estimation of registration rate at the Osaka Cancer Registry between 1966 and 1992 using this index. *Japanese journal of public health*, 45(10), pp. 1011–1017.

Akhwale, W. Bochner, A. F. Kwach, J. Liku, N. Muthee, V. Odhiambo, J. Onyango, F. Osterman, A. Prachi, M. Puttkammer, N. Wamiche, J. 2018. The impact of routine data quality assessments on electronic medical record data quality in Kenya. *PLoS ONE*, 13(4).

Amoroso, C. Basinga, P. Binagwaho, A. Gaju, E. Gashayija, M. Hedt-Gauthier, B. Hirschhorn, L. R. Iyer, H. S. Muhire, A. Nisingizwe, M. P. Rubyutsa, E. Wilson, R. 2014. Toward utilization of data for program management and evaluation: Quality assessment of five years of health management information system data in Rwanda. *Global Health Action*, 7(1), article no: 25829.

Anderka, M. Canfield, M. A. Copeland, G. Feldkamp, M. L. Kirby, R. S. Krikov, S. Isenburg, J. Mai, C. T. Mosley, B. Olney, R. S. Rickard, R. Romitti, P. A. Stanton, C. 2015. Development and implementation of the first national data quality standards for population-based birth defects surveillance programs in the United States Health policies, systems and management in high-income countries. *BMC Public Health*, 15(1), article no: 925.

Arboe, B. Christensen, J. H. Clausen, M. R. De Nully Brown, P. El-Galaly, T. C. Gørløv, J. S. Klausen, T. W. Munksgaard, P. S. Nygaard, M. K. Stoltenberg, D. 2016. The Danish National Lymphoma Registry: Coverage and Data Quality. *PLoS ONE*, 11(6).

Arts, D. G. T. De Keizer, N. F. Scheffer, G-J. 2002. Defining and improving data quality in medical registries: A literature review, case study, and generic framework. *Journal of the American Medical Informatics Association*, 9(6), pp. 600–611.

Asterkvist. A. Eloranta, S. Krawiec, K. Löfgren, L. Lönnqvist, C. Sandelin, K. 2019. Validation of data quality in the Swedish National Register for Breast Cancer. *BMC public health*, 19(1), article no: 495.

Ayatollahi, H. Khorasani-Zavareh, D. Mashoufi, M. 2019. Data Quality Assessment in Emergency Medical Services: What Are the Stakeholders' Perspectives? *Perspectives in health information management*, 16(Winter), article no: 1c.

Baesens, B. Dejaeger, K. Hamers, B. Poelmans, J. 2010. A Novel Approach to the Evaluation and Improvement of Data Quality in the Financial Sector. *Proceedings of the 15th International Conference on Information Quality*. Little Rock, USA. November 12–14, 2010.

Baesens, B. Dejaeger, K. Lemahieu, W. Moges, H-T. 2013. A multidimensional analysis of data quality for credit risk management: New insights and challenges. *Information & Management*, 50(1), pp. 43–58.

Bah, E. Hall, A. J. Shimakawa, Y. Wild, C. P. 2013. Evaluation of data quality at the Gambia national cancer registry. *International Journal of Cancer*, 132(3), pp. 658–665.

Ballou, D. P. Pazer, H. L. 1985. Modeling data and process quality in multi-input, multi-output information systems. *Management Science*, 31(2), pp. 150–162.

Ballou, D. P. and Tayi, G. K. 1998. Examining data quality. *Communications of the ACM*, 41(2), pp. 54–57.

Bank for International Settlements. 2011. Basel III: A global regulatory framework for more resilient banks and banking systems. Revised June 2011. Available at: <https://www.bis.org/publ/bcbs189.pdf>

Bank for International Settlements. 2013. Principles for effective risk data aggregation and risk reporting. Available at: <https://www.bis.org/publ/bcbs239.pdf>

Bank for International Settlements. 2019. Annual Report 2018/19. Promoting global monetary and financial stability. Available at: <https://www.bis.org/about/areport/areport2019.pdf>

Barker, P. M. Bennett, B. Mate, K. S. Mphatswe, W. Ngidi, H. Reddy, J. Rollins, N. 2012. Improving public health information: A data quality intervention in KwaZulu-Natal, South Africa. *Bulletin of the World Health Organization*, 90(3), pp. 176–182.

Batini, C. Scannapieca, M. 2006. *Data quality: Concepts, methodologies and techniques*. Springer-Verlag, Berlin, Heidelberg.

Batini, C. Scannapieco, M. 2016a. *Data and Information Quality: Dimensions, Principles and Techniques*. Springer International Publishing, Cham.

Batini, C. Scannapieca, M. 2016b. Introduction to Information Quality. Information Quality: The Potential of Data and Analytics to Generate Knowledge, pp. 1–17.

Blevins, J. Demyanenko, V. S. Fonarow, G. C. Hernandez, A. F. Olson, D. M. Peterson, E. D. Reeves, M. J. Schwamm, L. H. Smith, E. E. Webb, L. E. Xian, Y. Zhao, X. 2012. Data quality in the American Heart Association Get With The Guidelines-Stroke (GWTG-Stroke): Results from a National Data Validation Audit. *American Heart Journal*, 163(3), pp. 392–398.

Borek, A. Oberhofer, M. Parlikad, A. K. Woodall, P. 2011. A classification of data quality assessment methods. *Proceedings of the 16th International Conference on Information Quality*. Adelaide, Australia. November 18–20, 2011. pp. 189–203.

Borek, A. Parlikad, A. K. Woodall, P. 2013. Data quality assessment: The Hybrid Approach. *Information & Management*, 50(7), pp. 369–382.

Box, T. L. Byrd, J. B. Fihn, S. D. Maddox, T. M. Plomondon, M. E. Rumsfeld, J. S. Vigen, R. 2013. Data quality of an electronic health record tool to support VA cardiac catheterization laboratory quality improvement: The VA Clinical Assessment, Reporting, and Tracking System for Cath Labs (CART) program. *American Heart Journal*, 165(3), pp. 434–440.

Bray, F. Fedorenko, Z. Ferlay, J. Gorokh, Y. Goulak, L. Ryzhov, A. Soumkina, O. Znaor, A. 2018. Evaluation of data quality at the National Cancer Registry of Ukraine. *Cancer Epidemiology*, 53(1), pp. 156–165.

Bray, F. Johannesen, T. B. Langmark, F. Larsen, I. K. Møller, B. Parkin, D. M. Småstuen, M. 2009. Data quality at the Cancer Registry of Norway: An overview of comparability, completeness, validity and timeliness. *European Journal of Cancer*, 45(7), pp. 1218–1231.

Bray, F. Parkin, D. M. 2009a. Evaluation of data quality in the cancer registry: Principles and methods. Part I. Comparability, validity and timeliness. *European Journal of Cancer*, 45(5), 747–755.

Bray, F. Parkin, D. M. 2009b. Evaluation of data quality in the cancer registry: Principles and methods Part II. Completeness. *European Journal of Cancer*, 45(5), pp. 756–764.

Brenner, H. 1995. Use and limitations of the capture-recapture method in disease monitoring with two dependent sources. *Epidemiology*, 6(1), pp. 42–48.

Busacker, A. Harrist, A. Kroelinger, C. 2017. Evaluation of the Completeness, Data Quality, and Timeliness of Fetal Mortality Surveillance in Wyoming, 2006–2013. *Maternal and Child Health Journal*, 21(9), pp. 1808–1813.

Cappiello, C. Cerletti, C. Fratto, C. Pernici, B. 2018. Validating Data Quality Actions in Scoring Processes. *Journal of Data and Information Quality*, 9(2), pp. 1–27.

Charrondiere, U. R. Haytowitz, D. Nowak, V. Rittenschober, D. Stadlmayr, B. Wijesinha-Bettoni, R. 2016. Improving food composition data quality: Three new FAO/INFOODS guidelines on conversions, data evaluation and food matching. *Food Chemistry,* 193(1), pp. 75–81.

Chen, H. Jiang, L. Li, C. Ouyang, Y. 2015. A Multisource Retrospective Audit Method for Data Quality Optimization and Evaluation. *International Journal of Distributed Sensor Networks*, Special issue 2015, article no: 195015.

Clayton, H. Gulitz, E. Mahan, C. Petersen, D. Salihu, H. Sappenfield, W. Stanley, K. 2013. The Florida Investigation of Primary Late Preterm and Cesarean Delivery: The accuracy of the birth certificate and hospital discharge records. *Maternal and Child Health Journal*, 17(5), pp. 869–878.

Crocetti, E. Miccinesi, G. Paci, E. Zappa, M. 2001. An application of the two-source capture-recapture method to estimate the completeness of the Tuscany Cancer Registry, Italy. *European Journal of Cancer Prevention*, 10(5), pp. 417–423.

De Jongh, R. Joubert, M. Raubenheimer, H. Reynolds, E. Verster, T. 2017. A Critical Review Of The Basel Margin Of Conservatism Requirement In A Retail Credit Context. *The International Business & Economics Research Journal*, 16(4), pp. 257–274.

Eppler, M. J. Helfert, M. 2004. A classification and analysis of data quality costs. *Proceeding of the Ninth International Conference on Information* Quality. Cambridge, USA. November 5–7, 2004. pp. 311–325.

Espetvedt, M. Reksen, O. Rintakoski, S. Østerås, O. 2013. Data quality in the Norwegian dairy herd recording system: Agreement between the national database and disease recording on farm. *Journal of Dairy Science*, 96(4), pp. 2271–2282.

EU/575/2013. Regulation (EU) No 575/2013 on prudential requirements for credit institutions and investment firms and amending Regulation (EU) No 648/2012 Text with EEA relevance. European Parliament and Council. Brussels. 27.6.2013.

European Banking Authority. 2017. Guidelines on PD estimation, LGD estimation and the treatment of defaulted exposures. EBA/GL/2017/16. Available at: <https://eba.europa.eu/documents/10180/2033363/Guidelines+on+PD+and+LGD+estimation+%28EBA-GL-2017-16%29.pdf>

European Banking Authority. 2019a. Progress report on the IRB roadmap – monitoring implementation, reporting and transparency. Available at: <https://eba.europa.eu/documents/10180/2551996/Progress+report+on+IRB+roadmap.pdf>

European Banking Authority. 2019b. Missions and tasks [online]. Available at: <https://eba.europa.eu/about-us/missions-and-tasks> (Accessed: 4 September 2019)

European Central Bank. 2018a. ECB guide to internal models: Risk-type-specific chapters. Available at: <https://www.bankingsupervision.europa.eu/legalframework/publiccons/pdf/internal_models_risk_type_chapters/ssm.guide_to_internal_models_risk_type_chapters_201809.en.pdf>

European Central Bank. 2018b. Report on the Thematic Review on effective risk data aggregation and risk reporting. Available at: <https://www.bankingsupervision.europa.eu/ecb/pub/pdf/ssm.BCBS_239_report_201805.pdf>

European Central Bank. 2019. List of supervised entities. Available at: <https://www.bankingsupervision.europa.eu/ecb/pub/pdf/ssm.listofsupervisedentities201909.en.pdf.pdf>

European Union. 2019. Regulations, Directives and other acts [online]. Available at: <https://europa.eu/european-union/eu-law/legal-acts_en> (Accessed: 4 September 2019)

Ezell, J. D. Hazen, B. T. Jones-Farmer, L. 2014. Applying Control Chart Methods to Enhance Data Quality. *Technometrics*, 56(1), pp. 29–41.

Fisher, C. W. Lauria, E. J. M. Matheus, C. C. 2009. An Accuracy Metric: Percentages, Randomness, and Probabilities. *Journal of Data and Information Quality*, 1(3), pp. 1–21.

Funk, J. D. Lee, Y. W. Pipino, L. L. Wang, R. Y. 2006. *Journey to data quality*. MIT Press, Cambridge, Mass.

Granese, B. Gray, D. Heien, C. H. Joyce, H. I. Jugulum, R. Shi, C. Ramachandran, R. Singh, J. Talburt, J. R. 2015. Improving financial services data quality – a financial company practice. *International Journal of Lean Six Sigma*, 6(2), pp. 98–110.

Gray, C. L. Guidry, V. T. Hall, D. Lowman, A. Wing, S. 2015. Data quality from a longitudinal study of adolescent health at schools near industrial livestock facilities. *Annals of Epidemiology*, 25(7), pp. 532–538.

Gupta, S. Kulkarni, M. 2016. BCBS 239 Compliance: A Comprehensive Approach. *Cognizant 20-20 Insights*. Available at: <https://www.cognizant.com/whitepapers/bcbs-239-compliance-a-comprehensive-approach-codex1888.pdf>

Habibi, J. Mohsenzadeh, M. Vaziri, R. 2016. TBDQ: A Pragmatic Task-Based Method to Data Quality Assessment and Improvement. *PLoS ONE*, 11(5).

Heikkinen, S. Leinonen, M. K. Malila, N. Miettinen, J. Pitkäniemi, J. 2017. Quality measures of the population-based Finnish Cancer Registry indicate sound data quality for solid malignant tumours. *European Journal of Cancer*, 77(1), pp. 31–39.

Heinrich, B. Hristova, D. Klier, M. Schiller, A. Szubartowicz, M. 2018. Requirements for Data Quality Metrics. *Journal of Data and Information* Quality, 9(2), pp. 1–32.

Heinrich, B. Klier, M. 2011. Assessing data currency - a probabilistic approach. *Journal Of Information Science*, 37(1), pp. 86–100.

Hinterberger, H. Norrie, M. Presser, K. Weber, D. 2016. A scope classification of data quality requirements for food composition data. *Food Chemistry*, 193(1), pp. 166–172.

Holden, J. M. 1996. Expert systems for the evaluation of data quality for establishing the Recommended Dietary Allowances. *The Journal of nutrition*, 126(9S), pp. 2329–2336.

Jonasson, J. G. Jonsdottir, A. Olafsdottir, E. J. Olafsdottir, G. H. Sigurdardottir, L. G. Stefansdottir, S. Tryggvadottir, L. 2012. Data quality at the Icelandic Cancer Registry: Comparability, validity, timeliness and completeness. *Acta oncologica*, 51(7), pp. 880–889.

Lambe, M. Robinson, D. Sandin, F. Stattin, P. Tomic, K. Wigertz, A. 2015. Evaluation of data quality in the National Prostate Cancer Register of Sweden. *European Journal of Cancer,* 51(1), pp. 101–111.

Li, J. Wang, H. Yang, Z. Zhang, Y. 2014. Relative Accuracy Evaluation. *PLoS One*, 9(8).

Liaw, S-T. Taggart, J. Yu, H. 2015. Structured data quality reports to improve EHR data quality. *International Journal of Medical Informatics*, 84(12), pp. 1094–1098.

Lim, Y. M. F. Sivasampu, S. Yusof, M. 2018. Assessing primary care data quality. *International Journal of Health Care Quality Assurance*, 31(3), pp. 203–213.

Liu, Z. Pu, Y. Yin, K. Yu, Q. Zhou, B. 2014. An AHP-based Approach for Banking Data Quality Evaluation. *Information Technology Journal*, 13(8), pp. 1523–1531.

Majumdar, A. Ochieng, W. Y. Wilke, S. 2014. A framework for assessing the quality of aviation safety databases. *Safety Science*, 63(1), pp. 133–145.

Marsh, R. 2005. Drowning in dirty data? It’s time to sink or swin: A four-stage methodology for total data quality management. *Database Marketing & Customer Strategy Management*, 12(2), pp. 105–112.

Olson, J. E. 2003. *Data quality: The accuracy dimension*. Morgan Kaufmann publishers, San Francisco.

Prorokowski, H. Prorokowski, L. 2015. Solutions for risk data compliance under BCBS 239. *Journal of Investment Compliance*, 16(4), pp. 66–77.

Robert Morris Associates. 2016. Data quality – All roads lead to data. *The RMA Journal*, 98(9), pp. 46–51.

Robert Morris Associates. 2017. The 2016 RMA/AFS data quality survey – Are we there yet? *The RMA Journal*, 99(9), pp. 26–35.

Rutkowski, M. Tarca, S. 2016. Assessing the Basel II internal ratings-based approach. *Journal of Financial Regulation and Compliance,* 24(2), pp. 106–139.

Sadiq, S. Sharaf, M. A. Yeganeh, N. K. 2014. A framework for data quality aware query systems. *Information Systems*, 46(1), pp. 24–44.

Sebastian-Coleman, L. 2013. *Measuring data quality for ongoing improvement: A data quality assessment framework*. Morgan Kaufmann, Elsevier, Burlington.

Wang, R. Y. Strong, D. M. 1996. Beyond Accuracy: What Data Quality Means to Data Consumers. *Journal of Management Information Systems*, 12(4), pp. 5–33.

Weidema, B. P. Wesnaes, M. S. 1996. Data quality management for life cycle inventories – an example of using data quality indicators. *Journal of Cleaner* Production, 4(3–4), pp. 167–174.

Zhang, M. Peng, Y. Schuh, A. Megahed, F. M. Woodall, W. H. 2013. Geometric Charts With Estimated Control Limits. *Quality and Reliability Engineering International*, 29(2), pp. 209–223