data

MDPI

Article

Measuring the Effect of Fraud on Data-Quality Dimensions

Samiha Brahimi and Mariam Elhussein *

Department of Computer Information Systems, College of Computer Science and Information Technology, Imam Abdulrahman Bin Faisal University, P.O. Box 1982, Dammam 31441, Saudi Arabia
* Correspondence: maelhussein@iau.edu.sa

Abstract: Data preprocessing moves the data from raw to ready for analysis. Data resulting from fraud compromises the quality of the data and the resulting analysis. It can exist in datasets such that it goes undetected since it is included in the analysis. This study proposed a process for measuring the effect of fraudulent data during data preparation and its possible influence on quality. The five-step process begins with identifying the business rules related to the business process(s) affected by fraud and their associated quality dimensions. This is followed by measuring the business rules in the specified timeframe, detecting fraudulent data, cleaning them, and measuring their quality after cleaning. The process was implemented in the case of occupational fraud within a hospital context and the illegal issuance of underserved sick leave. The aim of the application is to identify the quality dimensions that are influenced by the injected fraudulent data and how these dimensions are affected. This study agrees with the existing literature and confirms its effects on timeliness, coherence, believability, and interpretability. However, this did not show any effect on consistency. Further studies are needed to arrive at a generalizable list of the quality dimensions that fraud can affect.

Keywords: data quality; quality dimensions; data analytics; preprocessing; fraud

Citation: Brahimi, S.; Elhussein, M. Measuring the Effect of Fraud on Data-Quality Dimensions. *Data* **2023**, *8*, 124. https://doi.org/10.3390/data8080124

Academic Editors: Pufeng Du and Ren-Hua Chung

Received: 7 March 2023
Revised: 13 April 2023
Accepted: 28 July 2023
Published: 30 July 2023

1. Introduction

Fraud can be defined as "An act of intentional deception or dishonesty perpetrated by one or more individuals, generally for financial gain" [1]. Another definition links fraud to personal gain [2]. The main characteristic of fraud is the effort to ensure it is undetected. When a human agent (i.e., an employee) interacts with the system with malicious intent, the stored data can be questioned. Fraud has many shapes and forms. Possible examples include credit cards, vendors, and payroll fraud. While there are known types of fraud, other types may exist within different organizations, contexts, and cultures. This depends on the benefits of the action to the individual committing fraud. Occupational fraud is the most challenging because employees know their internal systems [1]. They can make the data appear legitimate to ensure they are not exposed.

Research has explored various forms of fraud and fraud detection techniques. These studies aimed to uncover the actions and reveal those implicit using fraudulent data. However, the data resulting from fraud remains stored. This means that the quality of such data may be compromised, and the analysis results could be unreliable. While research recognizes many forms of fraud, very few studies have considered the impact of fraud activities on data quality dimensions [3].

Data Quality is an essential function of data management. Data Quality Management (DQM) involves employing processes, methods, and technologies to ensure that data quality meets specific business requirements [4]. Data preparation is one of the most important activities in data quality management. This may include data cleaning, transformation, and integration. During the lifecycle of data quality management, data quality is assessed against a set of business rules, and data preparation activities are planned accordingly.

In addition, it is necessary to identify the root cause of data quality issues to minimize repetitive data preparation activities. According to the Data Management Organization Guide [5], the root causes of data quality issues can be categorized into five classes: issues caused by lack of leadership, issues caused by data entry processes, issues caused by data processing functions, issues caused by system design, and issues caused by fixing issues. None of these categories included quality issues caused by fraudulent activity; a more recent categorization was provided by [6] in the context of healthcare. The authors classified data quality root causes into six categories borrowed from the Ishikawa diagram: issues caused by personnel capabilities, issues caused by materials, issues caused by machines, issues caused by methods, issues caused by management, and issues related to the mission. The latter includes two sub-categories related to healthcare insurance fraud: financial incentives or disincentives, and reimbursement systems. The authors mentioned the effect of medical insurance fraud, such as upcoding, on data quality.

This study proposes a process to measure the effect of fraudulent activities on data quality dimensions. The proposed model was applied to the case of undeserved, sick leaves.

2. Objectives

1. Propose a process to measure the effects of fraudulent data on data-quality dimensions.
2. Apply the proposed model to the case of undeserved sick leaves to investigate the affected dimensions by fraudulent data.
3. Identify patterns of fraudulent data effect on data quality dimensions

The remainder of the paper is organized as follows: Section 2 presents the literature review, Section 3 explains the research methodology, Section 4 presents the case of undeserved sick leaves, and Section 5 concludes the paper.

3. Background

3.1. Data Preparation and Data Quality

Data preparation is a major phase in data analysis. It consists of four main processes: data profiling, cleaning, integration, and transformation [7]. According to [8], data profiling evaluates data quality before performing data-cleansing activities. Data cleaning, one of the main processes of data preparation, was defined in [8] as improving data quality. The authors in [9] suggested using a data quality framework developed for relational databases as guidelines for performing data cleaning on electronic health records (EHR). They applied the proposed method to two types of EHR, and the results were promising. In contrast, the effect of imputing missing data on data quality was evaluated in [10]. Data integration combines data from multiple heterogeneous sources in a unified view to satisfy user queries [11]. The quality of the query response depends heavily on the data quality of the source systems. Thus, the authors of [11] suggested a framework for data integration using data quality. The framework evaluates the data sources using data quality measures and then ranks them using the top-K queries to select the highly ranked ones. Finally, data transformation is the process of transferring data from one format or structure to another. Aiming at improving FHIR (Fast Healthcare Interoperability Resources (FHIR) data quality, the authors in [12] developed a Java-based FHIR RDF data transformation toolkit to facilitate the use and validation of FHIR RDF (Resource Description Framework (RDF)) data. The present research functions in two data preparation processes: (1) data profiling for quality measurement and (2) data cleaning for fraudulent data removal (labelling).

3.2. Data Quality Assessment Methodology

According to [13], the conventional data-quality assessment consists of five main steps.

Data analysis: analyzing schema and other information (metadata) to fully understand the data and management rules.

DQ requirement analysis: Surveying the opinions of data users and administrators to identify quality issues and set new quality targets.

Identifying the critical area: The most relevant databases and dataflows were selected.

Process modeling: This provides a model of the processes that produce or update data.

Quality measurement: Select the quality dimensions affected by the quality issues identified in the DQ requirements analysis step and define corresponding metrics; measurement can be objective when it is based on quantitative metrics or subjective when it is based on qualitative evaluations by data administrators and users.

The research at hand is concerned with only some of the data quality assessment lifecycle, but only with steps contributing to the comparison between fraudulent and non-fraudulent data. These steps include identifying critical data, process modeling, and quality measurement. Details are provided in the methodology section.

4. Literature Review

The work in [14] determined inter-hospital differences in acute myocardial infarction (AMI-CFR), aiming to evaluate the extent to which Belgian discharge records allow the assessment of the quality of care in the field of AMI to identify the starting points for quality improvement. Sensitivity analysis was used to compare the collected datasets.

As part of a wider research work, [15,16] investigated the effect of introducing diagnosis-related group systems (DRG) in 2004 on the distribution of admission weights in very low birth weight infants. A significant decrease in the number of admitted low-weight infants was observed in both studies. In Ref. [6], the change was linked to the control imposed by the new coding system. They questioned the quality of the data before the introduction of the DRG systems. However, these studies could have measured the quality dimensions and their associated business rules.

Authors in [17] studied the accuracy and validity of thyroid surgery administrative hospital data by comparing its measures to medical data measures. The authors observed important discrepancies that affected data quality. These discrepancies can be attributed to upcoding practices [6]. However, the authors did not link these dimensions to specific business rules (or associated metrics).

Finally, the authors in [3] discussed the effects of fraud on the data quality dimensions. It has been claimed that fraud in transactional systems may affect five dimensions: consistency, coherence, believability, timeliness, and interpretability. However, the authors did not provide objective evidence for their claims.

All the mentioned works (summarized in Table 1 compares related studies based on the use of data quality dimensions) consider data quality from a narrowed point of view; that is, the effect on quality dimensions needs to be studied more in detail using business rules and associated metrics. In addition, it has been observed that the case of undeserved sick leave, a healthcare occupational fraud, is not covered from this perspective.

Table 1. Compares related studies based on data quality dimensions.

Reference	Quality Dimensions (s) Considered	Quality Dimensions Measured?	Business Rules Used?	Objective/Subjective
[14]	None	No	No	Objective
[15]	None	No	Yes	Objective
[16]	None	No	Yes	Objective
[17]	Accuracy, validity	Yes	No	Objective
[3]	Consistency, coherence, believability, timeliness, and interpretability	No	No	Subjective

5. Methodology

A six-step process is proposed to assess which quality dimensions are affected by fraud data. Figure 1 shows these steps, and further explanation follows.

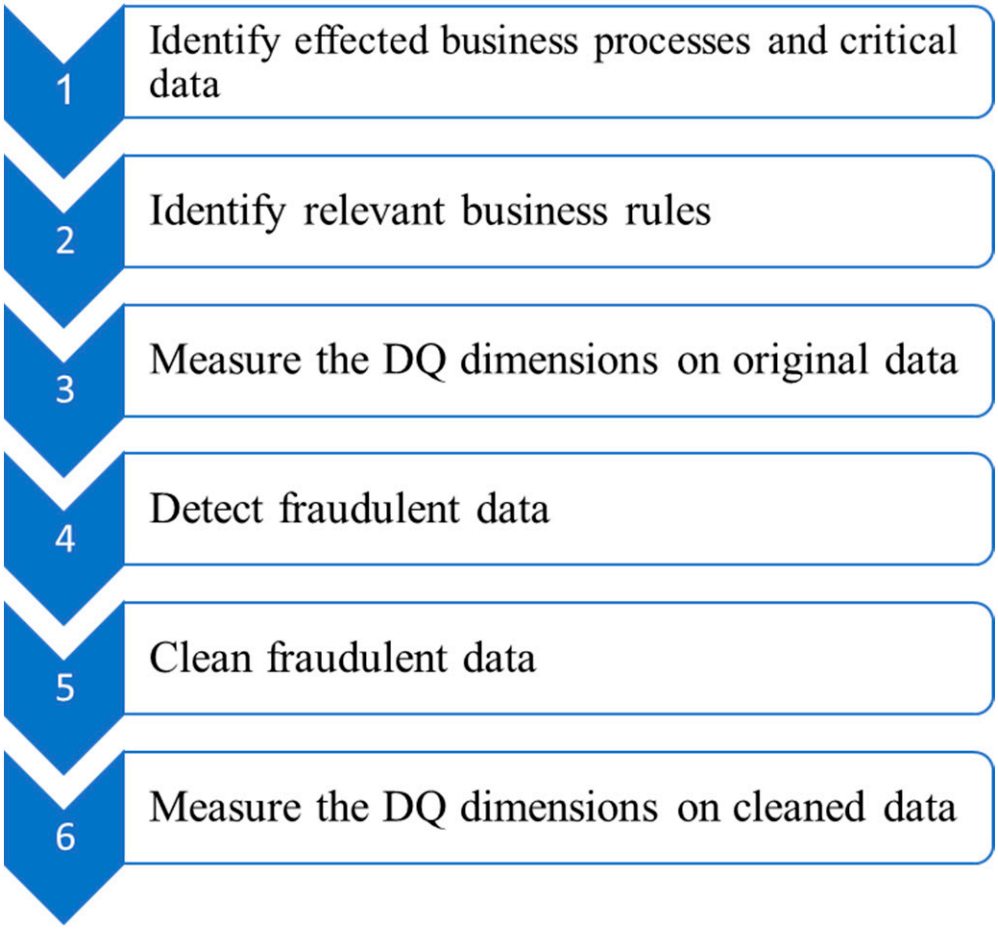

Figure 1. Steps of Measuring Fraud Effect of DQ.

A domain expert conducted the first step. He determines the business process affected by the fraudulent activity and the data they produce or update. These data will be evaluated. Subsequently, the organizational data quality business rules related to the business process in question were identified along with their associated DQ dimensions. Following this step, each business rule is measured on the original data, that is, the data before fraud is detected. Next, fraudulent data were detected. This may be as naïve as a subjective manual annotation. Alternatively, it can be performed using sophisticated machine learning and artificial intelligence techniques such as classification, clustering, and anomaly detection. In data cleaning, the identified data are treated either by removing fraudulent records/highlighting them or by correcting fraudulent values. A fraudulent value would make part of the entity (a record) illegitimate, but the entity itself still exists and is valid. However, fraudulent records result from an entirely illegitimate or non-existing entity. In the next step, the business rules are measured again, considering genuine data only: data that are not labeled as fraudulent, kept if fraudulent data are removed, or data with correct values.

6. Analysis and Results

The data were obtained from a local hospital, where some records were known to have been inserted to obtain undeserved sick leaves. These are sick leave days prescribed to patients based on non-existing medical conditions. Such sick leaves result from an alleged fraudulent activity in which a physician issues it to a patient to take days off without a medical need. This practice is known here, and the country tries to combat it by imposing punishments and legal consequences.

7. Overview of the Case Study

Sick leaves are days off given to employees or students to recover from their medical conditions. The abuse of this service is called undeserved sick leave and is considered occupational fraud [18]. Undeserved sick leaves can be obtained in four ways:

1. Providing fake certificates with false stamps and signature
2. Providing fake certificates with true stamps and false signatures; such notes are generally certified by hospital staff such as nurses.
3. Pretending sickness symptoms and lying to doctors to obtain a correct certificate.
4. Providing a fully correct document issued by a physician.

Researchers have begun to investigate the problem of underserved sick leaves. In [19], a set of methods for analyzing individual changes in sick leave diagnoses overtime was discussed. The authors of [20] described sick leave patterns in Saudi Arabia based on data. A set of machine learning models for detecting underserved sick leaves in hospitals in Saudi Arabia was proposed [18]. Finally, a clustering model for feature selection applied to uncover undeserved sick leave sellers on social media was proposed in [21].

The dataset consists of 20,021 records of sick leaves entered by physicians. Each record shows information about the patient, treating physician, and sick leave. The data show the number of days of sick leave, the start date, and the diagnosis. It provides the MRN of patients, their sex, and their age. The dataset contained 17 attributes, some of which were removed during the initial cleaning and preparation of the data. At this stage, an evaluation of (Data Quality) DQ dimensions is required to demonstrate the effect of fraudulent data on data quality. Table 2 summarizes the dataset's attributes, their types, and descriptions after standard data cleaning.

Table 2. A summary of the attributes in the dataset.

Attribute	Type	Description
ID	Numerical	Identify instances
MRN	Numerical	Identify patients
Sex	Binominal	Patient gender male/female
Age	Numerical	Patient age
NAT	String	Patient nationality
Problem	String	Explaining the patient's complaint
Date	Date	The date of the patient's visit to the hospital
Employee code	Numerical	The ID of the physician who orders the sick leave
Department	String	The department where the physician belongs
Number of days	Numerical	The number of sick leave days ordered

Owing to the problem of undeserved sick leaves, there is a need for a domain expert to assist with the identification process. The expert created two additional attributes to evaluate each record: Sp_match and Dx_match. Sp_match identifies the matching between the physician's specialization and the problem for which the patient was granted sick leave. Dx_match examines the match between a patient's complaint and the number of days given as sick leave. Both attributes take one of the following values: 1 if there is a match, 0 if there is a mismatch, or two if the expert cannot decide. Table 3 provides a statistical description of this dataset. Following this introduction, the proposed methodology was applied.

Table 3. Statistical description of the dataset.

Attribute	Count
MRNs	12,759
Sex	Almost 1:1 ratio of males and females
Age	They range between 1–88 years.
NAT	There are locals and non-locals; however 91% of the dataset is locals.
Problem	Textual description of the problem. The text is of various lengths.
Date	The dates span between March 2017–April 2018
Employee code	There are a total of 736 physicians in the dataset.
Departments	There are a total of 21 departments.
Number of days	The average number of days given is 1.6. 47% of the given sick leaves were one days and 30% were 2 days.

7.1. Identify Affected Business Processes and Critical Data

Per the expert, the business process (and its produced and updated data) affected by issuing undeserved sick leaves is "issuing a sick leave." Hence, sick leave records were the subject of this study.

7.2. Identify Relevant Business Rules and DQ Dimensions

Eleven business rules were identified concerning the process of issuing sick leaves. Some of these business rules have already been mapped to data-quality dimensions, whereas others still need to be mapped to data-quality dimensions. A subject area expert was consulted to support mapping the rest of the business rules with data quality dimensions. The list of dimensions studied was limited to the basic dimensions mentioned in [5], and those judged to be affected by fraudulent data [3]. The following section discusses and maps these business rules into seven dimensions.

Accuracy refers to the degree to which data represents real-life entities. A sick leave in the system is said to be accurate if it is represented by a sick leave document delivered to the patient. This is described by business rule BR1 in Table 4.

Completeness refers to whether all required data are present. Although most of the fields in the dataset should be complete, only two business rules related to issuing sick leaves were identified. BR2 and BR3 in Table 4 concern the population of the fields "Problem" and "Number of days," respectively. The reason for not covering the rest of the fields is that their completeness is ensured through system design.

Timeliness refers to the chronological patterns of transactions commonly observed in the data and information exchanges between subsystems. This was interpreted as the total number of sick leaves obtained repeatedly (day/week/month). A stable pattern reflects high timeliness. See BR4 in Table 4.

According to [5], uniqueness states that no entity exists more than once in a dataset. In this case, a business rule is identified in relation to uniqueness. It states that a patient can take only one sick leave at a time. See Table 4, BR5.

Validity refers to whether data are consistent with a defined domain of values. Considering the process of issuing sick leave, the field "Problem" is said to be valid if its value is among a set of problems defined by the hospital. For example, vitamin D deficiency is a problem that requires medical consultation but does not require sick leave. In Table 4, is defined as BR6.

Table 4. Business rules for the Undeserved Sick Leaves Case.

Dimension	Business Rule	Measure	Metric	Status Indicator
Accuracy	BR1: All sick leaves in the system should have a correspondent document delivered to the patient	Compare the number of documents printed with the number of sick leave records	Calculate the percentage of issued documents with corresponding record compared with the overall number of issued sick leaves	The higher the percentage, the higher the accuracy
Completeness	BR2: The "problem" field should be populated	Compare the number of non-null values in the "problem" field	Percentage of non-null values in the "problem" field to the overall number of records	The higher the percentage, the higher is completeness
	BR3: The "Number of days" field should be populated	Compare the number of non-null values in the "Number of days" field	Percentage of non-null values in the "Number of days" field to the overall number of records	The higher the percentage, the higher is completeness
	BR4: The "Start date" field should be populated	Compare the number of non-null values in the "Start date" fields	Percentage of non-null values in the "Start date" field to the overall number of records	The higher the percentage, the higher the completeness.
Timeliness	BR5: Time patterns of issuing sick leaves should be stable	Compare the number of records issued daily	Calculate the standard deviation of the number of issued sick leaves daily	A lower standard deviation means higher timeliness
Uniqueness	BR6: The patient cannot have more than one sick leave on the same day: {MRNs, start date} should be unique	Compare the number of unique to the overall records	Calculate the percentage of unique to the overall all number of records	The higher the percentage, the higher the uniqueness
Validity	BR7: the problem for which the patient has had a sick leave should be within a specific domain	Compare the valid records with the overall records	Calculate the percentage of valid records to the overall all number of records	The higher the percentage, the higher the validity
Coherence	BR8: The diagnosis of the patient and the physician's specialization should match	Compare the number of matching records with the non-matching records	Calculate the percentage of matching to the overall records	The highier the percentage, the higher the coherence
Believability: Consistency with rational and organizational rules	BR9: The number of days of sick leave should not exceed the number of days specified in the sick leaves guide	For each record, compare the number of issued days respecting the max number for the given diagnosis. Then, count the number of records that have a mismatch and compare it with the overall number of records	Calculate the percentage of matching to the overall records	The higher the percentage, the higher the believability
	BR8: The diagnosis of the patient and the physician's specialization should match	Compare the number of matching records with the non-matching records	Calculate the percentage of non-matching to the overall records	The lower the percentage, the higher the believability
Believability: Source reliability	BR10: The number of sick leaves issued by the doctor should not be higher than a given threshold	The proportion of physicians adhering to the threshold.	percentage of physicians not exceeding the threshold of the overall number of active physicians in the hospital	The higher the percentage the higher the believability
Interpretability	BR11: Physicians should submit a full description of the patient's problem.	Availability of a detailed description of the patient's problem	Percentage of sick leaves associated with a full description of the problem to the overall number of issued sick leaves	The higher is the percentage the better is the interpretability

The authors in [3] define coherence as the agreement of the relationships between information streams. An existing business rule related directly to coherence is the match between the diagnosis of the patient and the specialization of the treating physician (issuing sick leave). This business rule and its measures, metrics, and status indicators are described in Table 4 as in BR7. Believability: The extent to which the data are accepted in a specific environment and in accordance with relevant rules as true or as an item that seems true, real, and credible [22]. The starting point for detecting fraud from data is the unbelievability of the data. Believability is highly subjective, yet an attempt to measure it is presented in [22]. The authors suggested using four other dimensions to measure believability: accuracy, consistency with rational and organizational rules, resource appropriateness, and consistency with previous experiments. The accuracy was calculated objectively by comparing the data to the real world. As mentioned before, once sick leave is issued by a doctor, it is printed and delivered to the patient, which means that accuracy regarding sick leave is always fully met. Consistency with rational and organizational rules was measured by an expert on a scale of 0 to 1. The role of the expert is to specify the rational and organizational rules that should be considered. The expert in the casestudy has specified two main rules:(1) the match between the diagnosis and the physician's specialization; (2) the number of days should be consistent with the diagnosis; for instance, the flu does not need more than two days. Resource appropriateness is calculated using two measures: relevance and reliability. Because sick leaves can only be issued by doctors, their relevance is fully ensured. However, its reliability remains questionable. The expert mapped the physician's reliability to the frequency of sick leaves being issued. Finally, the expert excluded the "Previous Experiments in comparison" dimensions, as no previous experiments were available. Consequently, the weights of accuracy and consistency with previous experience are zero, which excludes them from the equation. The authors in [22] used all aspects discussed to calculate one value of believability. However, it is not possible to identify the weight of each related dimension, particularly when the subjectivity level is very high. Hence, in the case at hand, business rules are mapped for each subdimension separately. See BR8, BR9, and BR10 in Table 4. Interpretability: This represents explaining the system traces left by operators' activities. This permits us to explain the pertinence of suspect imbalances, which appear when one or more transaction elements are disregarded or when anomalous traces are found. In the context of the case at hand, this was seen as a patient problem description (see BR11 in Table 4).

7.3. Measuring DQ Dimension on Original Data

DQ dimensions were measured based on the metrics listed in Table 4. However, further interpretation is needed for how each metric is calculated based on the data in the case study to further explain the obtained results. Accuracy: The system does not allow sick leave to be printed unless the record is successfully saved in the database. This means that the database represents all records that have been printed, and accuracy is not suspected in this context.

Completeness: Three business rules were identified for this DQ dimension concerning the problem, number of days, and start date attributes. To calculate each business rule, the number of sick leaves issued with the "problem" attribute holding a null value was counted. Subsequently, the percentage of this count is calculated based on the total number of sick leaves issued. The process is repeated for the number of days and start date attributes.

Timeliness: A pivot table combining all sick leaves issued daily was constructed. The standard deviation of the grouped number of sick leaves was then calculated.

Uniqueness: The number of duplicate pairs of MRN (start date) was calculated to determine how often one patient received more than one sick leave on the same day.

Validity: To calculate the business rule related to validity, a pivot table showing all problems inserted by physicians to justify a sick leave issue was created. Each problem was evaluated subjectively against the conditions that entailed sick leave based on hospital

guidelines. An example of an invalid problem would be giving sick leave to someone complaining of a vitamin D deficiency.

Coherence: The number of sick leaves for which physician specialization does not match the diagnosis of the patient that s/he issued the sick leave is calculated. This is provided using the SP_mismatch attribute in the dataset. An example of a non-match is when an (ear, nose, and throat) ENT specialist issues a sick leave with an abdominal pain diagnosis. Believability is measured through consistency and source reliability.

To calculate the effect on consistency with rational and organizational rules, there are two metrics: the percentage of non-matching between the diagnosis and the number of days given in sick leave. This was provided as the percentage of non-matching sick leaves to the total number of sick leaves. The expert domain also provided a non-match between the diagnosis and the number of days given. There are guidelines for the maximum allowed sick leave days according to the severity of the diagnosis and the need to rest. An example of such a mismatch is to give four days off for an "acute upper respiratory infection", also known as flu. Another metric is calculating the maximum number of days given for each diagnosis and then counting and comparing the number of sick leaves within this maximum threshold. Only diagnoses that appeared > 20 times were included in this metric. Many diagnoses appear infrequently. For ease of calculation, only those that appeared frequently are included. The number of times these sick leaves were issued was compared with the maximum number of days for such a diagnosis. Subsequently, they were compared to the Dx_match attribute. Sick leaves were counted as a mismatch. The percentage of sick leaves among the overall sick leaves was calculated.

Source reliability: Based on a specific threshold, the percentage of physicians who exceeded this threshold was calculated using this metric. This number represents the annual number of sick leaves issued by physicians. For demonstration, this was set up as 100 sick leaves per year.

The last dimension to consider is interpretability. Physicians should provide a full description of a patient's complaints. To measure interpretability, the number of words used in the diagnosis can be used to measure the level of detail provided by the physician. To compute this, a minimum of three words that describe the problem must be considered. The percentage of sick leaves with full description was calculated based on this threshold. Table 5 provides a summary of the results obtained based on each metric.

Table 5. Summarizes the results of each metric performed on the data before removing undeserved leaves.

Dimension		BR#	Metric	Value
Accuracy		BR1	Guaranteed	100%
Completeness	problem	BR2	Percentage of non-null values in the "problem" field to the overall number of records	100%
	number of days	BR3	Percentage of non-null values in the "number of days" field to the overall number of records	100%
	start date	BR4	Percentage of non-null values in the "start date" field to the overall number of records	21.14%
Timeliness		BR5	The standard deviation of daily sick leaves	41.55
Uniqueness		BR6	Percentage of unique to the overall all number of records	100%
Validity		BR7	Percentage of valid records to the overall all number of records	99.984%
Coherence		BR8	Percentage matching physician specialization/diagnosis	96.53%
Believability	Consistency with rational • organizational rules	BR9	Percentage of sick leaves that were assigned a maximum number of days and were marked to be not matching	99.19%
		BR8	Percentage of matching number of days/diagnosis	96.59%
	Source reliability	BR10	Percentage of physicians not exceeding 100 sick leave per year	96.74%
Interpretability		BR11	Percentage of sick leaves having at least three words in the description	80.99%

7.4. Data Cleaning

A machine learning model was applied to identify undeserved leaves. In a previous study, machine-learning models were developed [18]. The models built and tested Naive bayes (NB), K-Nearest Neighbors (KNN), and Logistic regression. It also considered the class imbalance problem, as undeserved sick leaves are expected to be much less than authentic non-fraudulent leaves. Four proportions of the dataset with different ratios among the classes (deserved Vs undeserved) have been created. Each classification technique is evaluated under the sampled data proportions considering a set of measures such as accuracy, specificity, and Area Under-Curve (AUC).

The LR classifier shows the best performance on the original data (accuracy = 97%, specificity = 76% and AUC = 87%), followed by NB, then K-NN. However, on the sampled data, NB outperformed both LR and K-NN with an accuracy of 90%, specificity of up to 94%, and AUC of up to 88%.

The Naïve Bayes model built on sampled data (34% deserved sick leaves and 66% undeserved sick leaves) has been deployed and applied to the dataset considered in this research. The dataset used in this stage is collected in a different timeframe than the dataset used for building the model, yet they both have the same structure and characteristics. The model labeled each record as undeserved or deserved. The model could identify 1075 undeserved records, representing 7% of the dataset. Although the cleaning process is an integral part of this work, it is beyond the scope of this paper to explain the details of the model used. The authors would like to stress that other detection methods can also be used. However, the model developed in another study using the same data was applied here for convenience.

The dataset was cleaned, and undeserved sick leaves were removed (not considered for further analysis) to measure the DQ dimensions without these records.

7.5. Measuring DQ Dimensions on Cleaned Data

After cleaning, the same metrics were used to calculate data. Table 6 summarizes the results after removing undeserved sick leaves from the dataset.

Table 6. Summaries of DQs before and after removing fraud data.

Dimension		Before Cleaning	After Cleaning	Best Possible Value after Cleaning	Change	Effect
Completeness-start date		21.14%	19.88%	22.34%	Decreased	Decreased completeness
Timeliness		9.45	8.92	8.06	Decreased	Better timeliness
Validity		99.984%	99.985%	100%	Increased	Better validity
Coherence		96.59%	98.79%	100%	Increased	Better coherence
Believability	Consistency with rational organizational rules	99.19% 96.59%	99.55% 98.79%	100% 100%	Increased Increased	Higher rational and organization rules
	Source reliability	96.74%	96.1%	100%	Decreased	Decreased source reliability
Interpretability		80.99%	87.12%	100%	Increased	Higher interpretability

8. Discussion

It has been observed that the completeness of the field start date has decreased. There were 15,788 records with null values on the start date. After cleaning, this number was reduced to 15,178. The number of records with null values cleaned was 610 of the 1075 records cleaned (56.67%). Although this proportion was considered significant, the percentage of completeness decreased as the total number of records decreased. It is

important to note that the best completeness value that can be obtained after cleaning this dataset is 22.34%, appearing when all undeserved sick leaves records cleaned have their start date unpopulated. This means that fraudulent data can negatively affect completeness, but this is not the case at hand. However, this case shows that fraudulent data creates a false picture of the quality of the dataset. That is, the injected complete data, in this case, slightly increased completeness. However, it may increase to an acceptable level in other cases, particularly if the metric value is close to the acceptable boundaries.

As mentioned, timeliness was measured using the standard deviation of the daily aggregate of sick leaves obtained from the hospital. The value obtained before cleaning the undeserved leaves is 9.45. The daily aggregate varied between 0 and 48 (see the daily aggregate histogram in Figure 2). After cleaning, the measure decreased to 8.92, and the daily aggregated varied between 0 and 44 (see Figure 3). The best value that may be obtained after cleaning is 8.06, which can be achieved if all undeserved sick leaves are issued during days with high aggregates, reducing the maximum number of days issued daily to 26 instead of 48. In other words, the effect of fraudulent data on timeliness was significant, as it showed a significant improvement after cleaning.

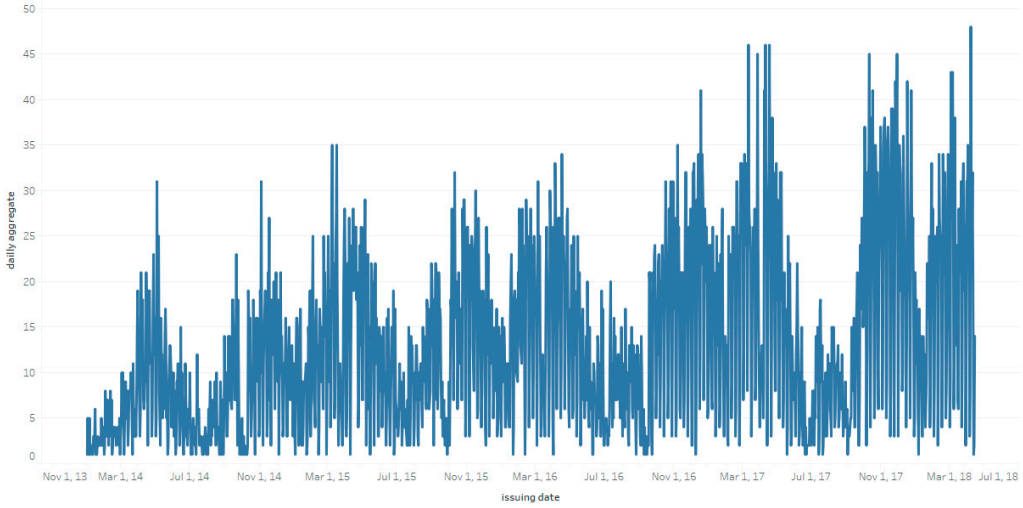

Figure 2. Daily aggregates histogram before cleaning.

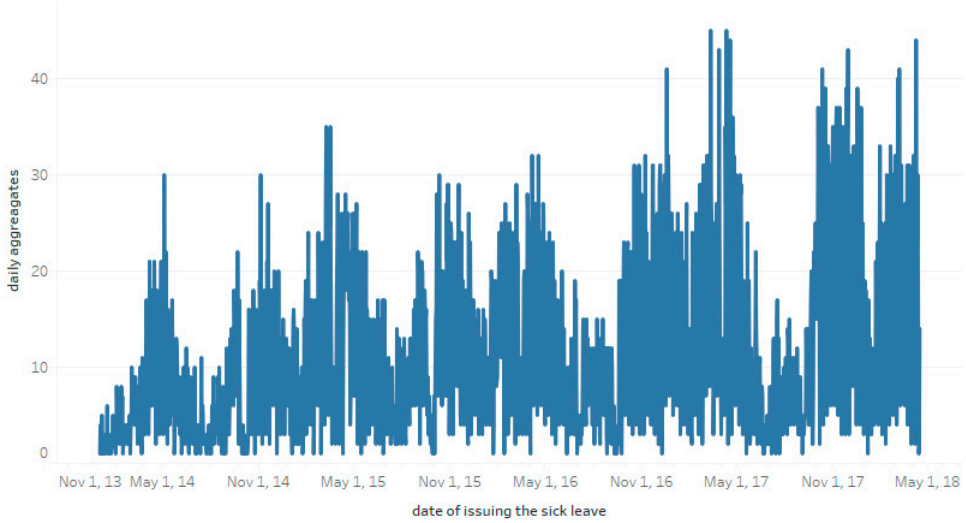

Figure 3. Daily aggregates histogram after cleaning.

Its validity before cleaning was high (99.984%). Only three records with invalid data were included in the dataset. After cleaning, none of the records were cleaned, as none were fraudulent. However, the validity after cleaning improved by 0.001 as the number of

records was reduced. This implies that the injection of fraudulent records affects validity or any similar dimension measured using the number of records.

Coherence is reflected by a match between physician specialization and the issue of sick leave. Before cleaning, coherence was 96.59%, which increased to 98.79% after cleaning. Overall, 683 records with physician specializations did not match the submitted diagnosis, among which only 23 records remained after cleaning. This demonstrates the significant effect of fraudulent data on this dimension.

Believability is one of the most complex, subjective dimensions. This is observed in the case at hand from two perspectives.

First, consistency with rational and organizational rules: this is expressed by two business rules. BR9 limits the number of days to those specified in the internal hospital guide. Before cleaning, the metric value of this business rule increased from 99.19% to 99.55%. Thus, 164 records with the maximum number of days were identified, and 85 records remained after cleaning. The second rule is BR8, which is mapped to coherence. The same effect has been reported previously.

Second, source reliability: physicians issuing a reasonable number (100 per year) of sick leaves per month (BR10). This evaluation has been performed in recent years. Before cleaning, 3.26% of physicians exceeded the threshold. In total, 16 doctors were 490 doctors. After cleaning, the number of doctors exceeded the threshold value (16). However, this percentage increased as the number of doctors decreased (from 490 to 410). The data of 80 doctors were removed during cleaning because all their issued sick leaves were undeserved. This shows that the business rule does not control fraudulent activity but is still affected by fraudulent data.

Interpretability is reflected in the availability of a detailed description of the case in which sick leave is issued (BR11). Before cleaning, 80.99% of the records were interpreted as being associated with the full case description. This percentage improved to 87.12% after cleaning fraudulent data. This may be explained by the inability to enter details about non-genuine diagnoses. Thus, interpretability is significantly affected by fraudulent data.

9. Conclusions, Implications, and Recommendations

This study proposes a process for evaluating the effect of fraudulent data on the data quality dimension during the preparation phase. The suggested model was applied to the occupational fraud case of undeserved sick leaves. The results reveal that fraudulent data can affect many dimensions in several ways. Some business rules are directly associated with fraud activity, others are not associated with it, and others are partially associated with it.

Business rules are directly associated with fraud activity, meaning there is a critical overlap between fraudulent data and data that do not adhere to the business rule. In this case, the metric value of the business rule should improve as fraudulent data that negatively affects business rules is removed. An example of this BR8 is mapped to coherence and believability (mismatch between diagnosis and physician specialization).

A business rule is not associated with fraudulent activity, which means that there is no overlap between fraudulent data and data not adhering to the business rule. However, the business rule measure is calculated based on the overall number of records. In this case, a lower number of records affected the metric value. For instance, the validity in the studied case (BR3) is measured using the percentage of invalid values in several fields among which the "problem". Only three records were identified as invalid and not part of the fraudulent data. However, the metric value decreased because the overall number of records decreased. This case may include a situation in which the metric is calculated based on some aggregates rather than on the overall size of the data, such as the source reliability in the case studied (BR10).

Business rules are partially associated with fraudulent activity, which means that a partial overlap with fraudulent data and measure is based on the overall number of records. In this case, the effect of removing fraudulent data and the effect of removing records from

the dataset may be contradictory, and the metric value remains at the boundaries of the old value. For example, the completeness of the "start date" is decreased because the number of records is lowered. However, if all fraudulent data had unpopulated start date, the metric value would have been improved (the best possible value was significant).

As mentioned, the authors in [3] subjectively discussed the effects of fraud on timeliness, coherence, consistency, believability, and interpretability. This research objectively confirmed the effects on timeliness, coherence, believability, and interpretability. However, the data at hand did not provide sufficient evidence for consistency. This does not deny that fraudulent data might significantly affect this dimension. However, further applications of the proposed method to other datasets are required to investigate this dimension. Moreover, the effect is also observed in other dimensions, namely, completeness and validity.

A limitation of this study is that it only investigated fraudulent records. Other applications of the model are required to consider the different granularities of the data, such as fraud at the attribute value level.

Author Contributions: Conceptualization, S.B. and M.E.; methodology, S.B.; software, S.B. and M.E.; validation, M.E.; formal analysis, S.B.; investigation, S.B.; resources, M.E.; data curation, S.B. and M.E.; writing—original draft preparation, S.B. and M.E.; writing—review and editing, S.B. and M.E.; visualization, M.E.; supervision, S.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Review Board (or Ethics Committee) of Imam Abdulrahman Bin Faisal University (protocol code IRB-2017-09-195 and 6 February 2019).

Informed Consent Statement: Not applicable.

Data Availability Statement: Data is unavailable due to privacy or ethical restrictions.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Gee, S. *Fraud and Fraud Detection: A Data Analytics Approach*; Wiley: Weinheim, Germany, 2014. [CrossRef]
2. Knepper, D.; Fenske, C.; Nadolny, P.; Bedding, A.; Gribkova, E.; Polzer, J.; Neumann, J.; Wilson, B.; Benedict, J.; Lawton, A. Detecting Data Quality Issues in Clinical Trials: Current Practices and Recommendations. *Ther. Innov. Regul. Sci.* **2016**, *50*, 15–21. [CrossRef]
3. Puentes, J.; Laso, P.M.; Brosset, D.; Puentes, J.; Laso, P.M.; Brosset, D.; Challenge, T. The Challenge of Quality Evaluation in Fraud Detection. *HAL* **2018**, *10*, 1–5. [CrossRef]
4. Allen, M.; Cervo, D. *Multi-Domain Master Data Management*; Elsevier: Amsterdam, The Netherlands, 2015; ISBN 9780128008355.
5. DAMA-DMBOK. *The DAMA Guide to the Data Management Body of Knowledge*, 2nd ed.; Technics Publications, LLC: Denville, NJ, USA, 2015; ISBN 0977140083.
6. Carvalho, R.; Lobo, M.; Oliveira, M.; Raquel, A.; Alonso, V.; Lopes, F.; Souza, J. Analysis of Root Causes of Problems Affecting the Quality of Hospital Administrative Data: A Systematic Review and Ishikawa Diagram. *Int. J. Med. Inform.* **2022**, *156*, 104584. [CrossRef] [PubMed]
7. Jassim, M.A.; Abdulwahid, S.N. Data Mining Preparation: Process, Techniques and Major Issues in Data Analysis. *IOP Conf. Ser. Mater. Sci. Eng.* **2021**, *1090*, 012053. [CrossRef]
8. Ganti, V.; Samara, A. *Das Data Cleaning a Practical Perspective*; Morgan & Claypool Publishers: Kentfield, CA, USA, 2013; ISBN 9781608456772.
9. Dziadkowiec, O.; Callahan, T.; Ozkaynak, M.; Reeder, B.; Welton, J. Using a Data Quality Framework to Clean Data Extracted from the Electronic Health Record: A Case Study. *eGEMs* **2016**, *4*, 11. [CrossRef] [PubMed]
10. Philip, S.; Vashisth, P.; Chaturvedi, A.; Gupta, N. Data Quality Improvement by Imputation of Missing Values. In Proceedings of the 2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA), Coimbatore, India, 2–4 September 2021.
11. Abdel Monem, R.; Bastawissy, A.H.E. DIRA: A Framework of Data Integration Using Data Quality. *Int. J. Data Min. Knowl. Manag. Process* **2016**, *6*, 37–58. [CrossRef]
12. Prud'hommeaux, E.; Collins, J.; Booth, D.; Peterson, K.J.; Solbrig, H.R.; Jiang, G. Development of a FHIR RDF Data Transformation and Validation Framework and Its Evaluation. *J. Biomed. Inform.* **2021**, *117*, 103755. [CrossRef] [PubMed]

13. Batini, C.; Cappiello, C.; Francalanci, C.; Maurino, A. Methodologies for Data Quality Assessment and Improvement. *ACM Comput. Surv.* **2009**, *41*, 16. [CrossRef]
14. Aelvoet, W.; Terryn, N.; Molenberghs, G.; De Backer, G.; Vrints, C.; Van Sprundel, M. Do Inter-Hospital Comparisons of in-Hospital, Acute Myocardial Infarction Case-Fatality Rates Serve the Purpose of Fostering Quality Improvement? An Evaluative Study. *BMC Health Serv. Res.* **2010**, *10*, 334. [CrossRef]
15. Abler, S.; Verde, P.; Stannigel, H.; Mayatepek, E.; Hoehn, T. Effect of the Introduction of Diagnosis Related Group Systems on the Distribution of Admission Weights in Very Low Birthweight Infants. *Arch. Dis. Child. Fetal Neonatal Ed.* **2011**, *96*, F186–F189. [CrossRef]
16. Freitas, A.; Gaspar, J.; Rocha, N.; Marreiros, G.; Da Costa-Pereira, A. Quality in Hospital Administrative Databases. *Appl. Math. Inf. Sci.* **2014**, *8*, 1–6. [CrossRef]
17. Mercier, F.; Laplace, N.; Mitmaker, E.J.; Colin, C.; Kraimps, J.L.; Sebag, F.; Bourdy, S.; Duclos, A.; Lifante, J.C. Unexpected Discrepancies in Hospital Administrative Databases Can Impact the Accuracy of Monitoring Thyroid Surgery Outcomes in France. *PLoS ONE* **2018**, *13*, e0208416. [CrossRef]
18. Brahimi, S.; El Hussein, M.; Al-Reedy, A. Detection of Undeserved Sick Leaves in Hospitals Using Machine Learning Techniques. *Sustain. Comput. Inform. Syst.* **2022**, *35*, 100665. [CrossRef]
19. Hagberg, J.; Vaez, M.; Alexanderson, K. Methods for Analysing Individual Changes in Sick-Leave Diagnoses over Time. *J. Prev. Assess. Rehabil.* **2010**, *36*, 283–293. [CrossRef] [PubMed]
20. Elabd, K.; Alkhenizan, A.; Aldughaither, A. Sick Leaves Pattern in a Tertiary Healthcare Facility in Saudi Arabia. *Cureus* **2020**, *12*, e11543. [CrossRef] [PubMed]
21. Elhussein, M.; Brahimi, S. Clustering as Feature Selection Method in Spam Classification: Uncovering Sick-Leave Sellers. *Appl. Comput. Inform.* **2021**. [CrossRef]
22. Moossavizadeh, S.M.H.; Mohsenzadeh, M.; Arshadi, N. A New Approach to Measure Believability Dimension of Data Quality. *Manag. Sci. Lett.* **2012**, *2*, 2565–2570. [CrossRef]