applied system innovation

MDPI

Article

Data Governance in the Health Industry: Investigating Data Quality Dimensions within a Big Data Context

Suraj Juddoo [1,*], Carlisle George [2], Penny Duquenoy [2] and David Windridge [2]

1 School of Science and Technology, Middlesex University Mauritius, Uniciti, Cascavelle 90203, Mauritius
2 Department of Computer Science, School of Science and Technology, Middlesex University, London NW4 4BT, UK; c.george@mdx.ac.uk (C.G.); p.duquenoy@mdx.ac.uk (P.D.); d.windridge@mdx.ac.uk (D.W.)
* Correspondence: s.juddoo@mdx.ac.mu; Tel.: +230-57592191

Received: 30 September 2018; Accepted: 26 October 2018; Published: 1 November 2018

Abstract: In the health industry, the use of data (including Big Data) is of growing importance. The term 'Big Data' characterizes data by its *volume*, and also by its *velocity*, *variety*, and *veracity*. Big Data needs to have effective data governance, which includes measures to manage and control the use of data and to enhance data quality, availability, and integrity. The type and description of data quality can be expressed in terms of the dimensions of data quality. Well-known dimensions are *accuracy*, *completeness*, and *consistency*, amongst others. Since data quality depends on how the data is expected to be used, the most important data quality dimensions depend on the context of use and industry needs. There is a lack of current research focusing on data quality dimensions for Big Data within the health industry; this paper, therefore, investigates the most important data quality dimensions for Big Data within this context. An inner hermeneutic cycle research approach was used to review relevant literature related to data quality for big health datasets in a systematic way and to produce a list of the most important data quality dimensions. Based on a hierarchical framework for organizing data quality dimensions, the highest ranked category of dimensions was determined.

Keywords: Big Data; data quality; health data; data quality dimensions

1. Introduction

Big Data refers to the capacity to work with datasets using tools different to those used with traditional relational databases [1]. Big Data is generally characterized by *volume*, *variety*, and *velocity*. However, *veracity* is another characteristic of Big Data which is growing in popularity and concerns the rising issue of certainty or quality involved with the use of data. Data quality (DQ) is explained as data 'fit for use' [2]; this broad definition conveys the notion that data is used for certain objectives, and data of high quality would be data which is adequate enough to allow the users of data to meet their objectives. In the realm of Big Data, research on quality is still at an infancy stage, triggering the need for more research in this domain.

The healthcare sector is said to be a multi-trillion-dollar company in the making [3]. It is an example of an industry that makes use of a huge amount of data. Health data can be categorized differently, such as electronic health records, administrative data, claims data, disease registries, health surveys, and clinical trials data, amongst others [3]. Google Flu and Ebola forecast systems are two of the most well-known ways Big Data has been applied in the health industry [3]. There are other lesser known ways, such as:

1. Big Data is being used to improve decision making in the healthcare industry by increasing the potential of evidence-based medicine's (EBM's) "small data" [4]. EBM is defined as "the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients" [5]. Personalized decision support systems (PDSS) are enhancing personalized medicine or evidence-based medicine through big data analytics [6].
2. Healthcare frauds are very serious issues in many countries. Big Data, with the help of data mining combined with machine learning, can play a major role in fraud detection. Data mining can identify some fraud as soon as it happens and therefore positively increase prevention.
3. Big Data analytics is being applied with the aim of reducing patients' readmission numbers. Patient readmission is not only very expensive for hospitals, but the ratio of patients' dying after readmission is alarmingly high [7].
4. Big data is proving to be a very useful tool for medical research. As there are many very large medical datasets, such as the human genomic dataset, pharmaceutical companies are harnessing the power of Big Data analytics to discover new medicines and understand diseases.
5. Through the use of Internet of Things (IoT) in healthcare, data is retrieved in a pervasive manner. The data collected through IoT governs the daily life of the patient. Through integrating Big Data and IoT with healthcare, both patients and health facilities cut down costs by reducing the repetition of tests, so they benefit from more accurate diagnoses.

Without proper data quality, most of the above use cases of Big Data in the healthcare industry will not be effective. Many authors involved with data quality argue that understanding the dimensions of data quality is the first step leading toward appropriate data quality activities [8–11]. Data quality dimensions are the means of expressing the notion of data quality; examples of very frequently cited data quality dimensions (DQDs) involved with Big Data are: consistency, accuracy, completeness, and timeliness [12]. In order to undertake any data quality initiative, it is imperative to ascertain the most important DQDs associated within the domain of data usage.

The principal goal of this paper is to discuss, analyze, and recommend DQDs suitable in the context of Big Data application in the health industry. Using the well-established hierarchical framework developed by [13], the paper also aims to confirm the importance and validity of each of the four main data quality dimension categories in the context stated above. Therefore, a major gap in knowledge would be addressed, as there is currently a lack of precise and scientific ways for evaluating the most important DQDs for Big Data within the health industry.

This paper first reviews relevant literature regarding DQDs in general, and specifically those adopted in the healthcare industry. Secondly, it discusses the application of an inner hermeneutic cycle (IHC) research approach to undergo a systematic review of literature related to data quality for big health datasets and to obtain a list of the most important DQDs. The main findings of the work are then be discussed and, finally, a summary of the results and their implications is given.

2. Literature Review

This section discusses the state of knowledge primarily related to DQDs. The link between data governance and DQ is first elaborated. Further, DQDs are discussed in general and within the specific contexts of Big Data and the healthcare industry.

2.1. Data Governance and Data Quality

Due to the growing awareness of the importance of using data correctly, many organizations are setting up data governance (DG) policies. Even if DG is often regarded as a pure information technology (IT) function, many authors believe that it should be an organization-wide concern [14]. The DG framework of the Data Governance Institute (DGI) defines DG as "the exercise of decision making and authority for data related matters" [15], and the principal benefits for organizations that

adopt DG are (1) increased revenue and value, (2) managed cost and complexity; and (3) ensured compliance, security, and privacy risk control.

The DGI data governance framework discusses DQ in terms of DG's focus on DQ [15]. This framework possesses six main focus areas, one of them being data quality. The idea is that DG programs aimed at improving DQ would normally involve the application of DQ software and could be applied locally to one department or throughout an organization/enterprise. Applying the data governance framework requires the involvement of data stewards and data stakeholders, and it is very often quite a challenging process. Other authors argue that having DG helps toward improving DQ [16]. They state that data governance officers can help improve the consistency and quality of data during the data life cycle instead of relying upon the IT department to achieve this enormous responsibility. The IBM Data Governance Council Maturity Model highlights data quality management as one of the core disciplines of effective data governance. It defines data quality management as "methods to measure, improve and certify the quality and integration of data" [17]. Other authors also list data quality as part of the decision domains for data governance [18]. Their logic is that DG should help set DQD-related standards, such as accuracy and completeness, but it should also create policies to facilitate the proper communication and evaluation of DQ measures.

The few research studies that have focused on Big Data governance also seem to indicate a clear link with DQ, e.g., [15,18,19]. Reviewing some of the major challenges and opportunities of Big Data, Malik [19] states the opinion that the *volume* and *velocity* of data will decrease the certainty associated with data. Furthermore, Malik claims that existing data quality technologies cannot cope with the uncertainty issues caused by the *volume* and *velocity* characteristics of Big Data.

2.2. Data Quality Dimensions

Throughout the literature, there have been different sets of DQDs considered by several authors. A brief discussion of some of the varied DQDs follows below.

Some authors have investigated how to measure or assess the level of the quality of data. They have argued that some assessments of data quality could be task-independent and, therefore, not restrained by the context of application, while others are task-dependent [20]. Table 1 depicts the main DQDs they thought were worthy of discussion, together with some accepted definitions.

Table 1. List of well-cited traditional data quality (DQ) dimensions [20].

Dimensions	Definitions
Accessibility	Extent to which data is available, or easily and quickly retrievable
Appropriate amount of data	Extent to which volume of data is appropriate for the task at hand
Believability	Extent to which data is regarded as true and credible
Completeness	Extent to which data is not missing and is of sufficient breadth and depth for the task at hand
Consistent representation	Extent to which data is presented in the same format
Ease of manipulation	Extent to which data is easy to manipulate and apply to different tasks
Free-of-error	Extent to which data is correct and reliable
Interpretability	Extent to which data is in the appropriate languages, symbols, and units, and the definitions are clear
Objectivity	Extent to which data is unbiased, unprejudiced, and impartial
Relevancy	Extent to which data is applicable and helpful for the task at hand
Reputation	Extent to which data is highly regarded in terms of its source and content
Security	Extent to which access to data is restricted appropriately to maintain its security
Timeliness	Extent to which data is sufficiently up-to-date
Understandability	Extent to which data is easily comprehended
Value-added	Extent to which data is beneficial and provides advantages from its uses

There have been several other research studies undertaken which cite the above as potential DQDs [2,21]. *Timeliness* was investigated in the context of information manufacturing systems [21]. The aim was to assess the trade-off between maximizing *accuracy* and *completeness* versus improving

timeliness. The results show a direct positive relationship between *timeliness* and *accuracy* in the sense that data recently created has been found to be more accurate, and that completeness depends upon the input time of data. This type of research demonstrates that the context of data use, in this case, information manufacturing systems, is very important in determining relevant DQ dimensions.

Panahy et al. [2] also discussed the trade-off between *accuracy* and *timeliness* while aiming to investigate the dependencies between four different DQ dimensions, namely: *accuracy*, *completeness*, *consistency*, and *timeliness*. The authors created a questionnaire called the ACCTI (accuracy, completeness, consistency and timeliness, improvement process) framework to assess correlations between the four DQDs cited above. After the application of multivariate statistical tests, the use of Bartlett's test appeared to imply that the four dimensions denote a high level of dependency based on answers about different information systems. Hence, there seems to be some indication that the above four DQ dimensions are normally very relevant for different types of information systems, even if the research did not involve Big Data.

However, there seems to be a lack of uniformity and standardization in the terms used to describe DQDs [9]. Some ideas have been expressed by interchangeable terms: for example, some authors have described the idea of *accuracy* by use of the term *completeness* [22]. In other cases, other terms have been used to describe the same dimension of accuracy. *Accuracy* has been referred to as *precision* and also *semantic accuracy* [23]. Dong et al. [24] explained the notion of accuracy as 'true value' in the context of data fusion, which refers to a process of integrating data from different sources together while maintaining a standard of data quality.

2.3. Dimensions of Data Quality for Big Data

The high *volume* and *velocity* properties of Big Data make it more challenging to distinguish between clean and dirty data for further data analysis. Also, due to data coming from multiple sources, there is a need for a more effective method of increasing data quality through machine learning approaches [25]. Other authors argue that the importance of improving data quality for Big Data might not be high, since the amount of incorrect data is deemed to be negligible and hence will not affect the final outcome of data analysis [26]. The amount and impact of the erroneous or 'dirty' data as part of a big dataset is crucial in determining the importance of data quality for Big Data. A better understanding of DQDs that is more relevant for Big Data will support research and industry in building more appropriate data quality tools.

Caballero et al. [23] posit that the main DQD to be considered for Big Data is *consistency*, which they explain as the capability of information systems to ensure uniformity of datasets when data are being transferred across networks and systems. Their main hypothesis is that the business value of a dataset can only be estimated in its context of use. They further subdivide *consistency* into three subsequent parts, as discussed below and seen in Figure 1. Additionally, they connected many of the traditional data quality dimensions with the three consistency subdomains, as follows:

1. *Contextual consistency* refers to how far big datasets are used within the same domain of interest independently of data format, size, and velocity of the production of data. For the current research, the domain of interest refers to health data. Relevancy, credibility, ease of understanding, accuracy, and confidentiality are key DQDs for this type of consistency.
2. *Temporal consistency* conveys the idea that data needs to be understood in a consistent time slot such that the same data might not be comparable if it is from another time slot. Time concurrency, availability, and currency are deemed to be essential for temporal consistency.
3. *Operational consistency* brings in the operational influence of technology on the production and use of data. The sources of data could be more than one in Big Data scenarios; hence, operational consistency is crucial for ensuring the veracity of data. Availability, portability, precision, completeness, and traceability are considered the main connected dimensions here. Table 2 shows a mapping of the 3Vs of Big Data (volume, variety, velocity) to the 3Cs (contextual consistency, temporal consistency, operational consistency) of data quality.

Table 2. Matrix of 3Cs (contextual consistency, temporal consistency, operational consistency) relative to the 3Vs (volume, variety, velocity) [23].

	Velocity	**Volume**	**Variety**
Contextual	Consistency, Credibility, Confidentiality	Completeness, Credibility	Accuracy, Consistency, Understandability
Temporal	Consistency, credibility, Currentness, Availability	Availability	Consistency, Currentness, Compliance
Operational	Completeness, Accessibility, Efficiency, Traceability, Availability, Recoverability	Completeness, Accessibility, Efficiency, Availability, Recoverability	Accuracy, Compliance, Accessibility, Efficiency, Traceability, Availability, Recoverability, Precision

Other research studies have focused on determining data quality dimensions for Big Data using the IHC research method [8]. However, the context of the research in [8] differs from the current work since the discussions were for Big Data in general, and not specific to the health industry domain. Furthermore, the authors in [8] use a different research hypothesis, focusing on three Big Data 'coordinates', namely: data types, sources, and application domains. They conducted research on specific types of data, such as maps, semi-structured texts, linked open data, sensor and sensor networks, and official statistics. Correlations between DQDs and the Big Data coordinates were reported as: *accuracy* for maps, *completeness* of official statistics, *readability* for semi-structured data, *accessibility* and *trust* for linked open data, and *consistency* for sensor and sensor networks. An IHC was carried out comprising an initial corpus of 1600 papers, related tables, and notes. Keywords as part of the titles and abstracts having a minimum thread of 100 citations for the period of 2005–2014 were used as criteria for sorting. A summary of the literature review results was used to devise their theoretical conceptual framework which detailed the Big Data quality dimensions clustered by the above-cited application areas, ranging from maps to official statistics.

The field of data stream management systems (DSMS) seems to be highly connected with Big Data, as both fields argue for being able to work with a large volume of data on a near real-time basis. The only difference lies in the variety characteristic of Big Data, and it is not clear whether it is also a requirement for DSMS [27]. Geisler et al. [27] conducted a research study to create an ontology-based data quality framework in that context. Their results are highly relevant for the work carried out in this paper since DSMS focus on the real-time aspect of data management systems, whereas most other work in the field of data quality tends to concentrate on the *volume* and *variety* aspects of Big Data quality. They quoted two categories of DQDs, namely: application-based and system-based DQDs. Thus, their hypothesis argues that data quality might vary between applications that are part of the same information system. Ultimately, they listed the following DQDs as the most important for DSMS: *completeness*, *volume*, *accuracy*, *timeliness*, *consistency*, and *confidence*. The distinguishing factor of their research is the inclusion of "volume" as a DQD, which is explained as the number of tuples or values that a result is based upon [27]. A final difference between DSMS and Big Data in the work of [27] is that their work was undertaken on tuples-based relational database systems, which might be quite different from non-relational techniques of Big Data systems.

2.4. Data Quality Dimensions Specific for the Healthcare Industry

The *veracity* characteristic of Big Data is argued to be of crucial importance for the healthcare industry [3]. The two main reasons forwarded are (i) inaccurate data could ultimately result in life or death decisions made with regard to patients and, therefore, *accuracy* is a very important dimension; (ii) the high level of incorrectness usually present in doctors' prescriptions leads to a lot of wastages and inefficiencies, at the very least. Hence, *correctness* could be derived to be another important dimension, according to the authors. Lastly, as the *veracity* characteristic refers to the confidence in the use of data, *believability* and *trust* could be assumed to be also extremely prominent DQDs.

In the context of using clinical data in the US healthcare industry, some industry-based whitepapers point toward the importance of the following six DQDs, namely: *timeliness*, *equitability*, *care-safe*, *patient-centered*, *effective*, and *efficient*. Most of the above-mentioned dimensions are not part of the traditional list of DQDs normally discussed in data quality literature, and thus strengthen the rationale for the need for further research in this area. The six key attributes (dimensions) proposed by the UK audit commission in 2007 to improve data quality are: *accuracy*, *validity*, *reliability*, *timeliness*, *relevance*, and *completeness*.

The Canadian Institute for Health Information (CIHI) Data Quality framework assessment tools updated in 2017 discusses the following groups of DQDs—*accuracy and reliability*, *timeliness and punctuality*, *comparability and coherence*, *accessibility and clarity*, and *relevance* [28].

The data quality framework of New Zealand Health Information Service (NZHIS) was adapted from the CIHI's data quality framework [28]. Furthermore, the NZHIS felt that there was a need for transparency in the quality dimensions, which brought the implantation of additional DQDs, such as privacy and security, by the ministry's senior advisors for the health sector. Those dimensions were generated basically to address privacy and security standards, legislation, policies, and processes, with the idea of ensuring the privacy and security of individual information. In conclusion, the DQDs given by NZHIS are *accuracy*, *timeliness*, *comparability*, *usability*, *relevance*, *privacy*, and *security.*

There have been previous research studies conducted with the aim of determining the level of accuracy in international disease coding under the ICD-9 (International Classification of Diseases, ninth revision) standard, specifically in New Zealand [29]. The major result is that 18% of the first four digits of E-codes and 8% of the location codes were incorrect. No major differences were noted between small and large hospitals. This work tends to confirm the perception that accuracy is one of the very important DQDs, even for the very specific type of data in the health industry. This research again was carried out outside the Big Data context. Other authors conducted a research study within the specific domain of trauma-related registries [30]. Their aims included the classification of DQDs and the investigation of measurement methods and of ways of improving DQ in general [30]. The point of interest in their work in relation to the research carried out in this paper is that they posited that the dimensions of *accuracy*, *completeness*, and *capture* are the most relevant DQDs to explore. The work in [30] also made use of an integrative literature review technique approach. The *capture* DQD seems to be different compared to the normal DQDs usually cited. The authors described the *capture* dimension as the extent to which all necessary cases that could have been registered have actually been registered. The existence of this *capture* dimension confirms the assumption that DQ is linked with its context of use, and, therefore, each different context of use might require different DQDs.

A recent study within the context of the Ebola outbreak in 2014 pointed to some interesting DQDs [31]. Even if the authors did not directly investigate DQDs, they discussed some of the major challenges to improving data quality for Ebola outbreaks. Focusing on the challenges, the authors cited *accuracy*, *privacy and security*, *heterogeneity*, *provenance and trust*, *availability*, and *completeness*. The authors advocated for the use of different tools for data capture, including websites and mobile apps. Some of the characteristics of Big Data seem to be present as part of the datasets examine; however, at no point in their work did the authors confirm whether they were using Big Data.

Almutiry et al. [32] proposed a framework for DQ in the context of cloud-based information systems. The authors performed a general literature review and then filtered out DQDs for redundancies. The process of filtering was not discussed and, therefore, the rationale for eliminating certain dimensions and how redundancies were defined in this context are not known. The result is a framework consisting of three sections: Information (*accuracy*, *completeness*, *consistency*, *relevance*, *timeliness*, and *usability*), Communication (*provenance*, *interpretability*), and Security.

Other authors have carried out research into DQDs in specific areas of health data. One such example is the work concerning DQDs for genome annotations [10]. The complexity and variety of genome annotations make it prone to suffer from DQ issues. The authors applied a scenario-based research method. Out of a total of 17 typical DQDs detected after the analysis of the scenarios, 5 main

DQDs were deduced after the statistical analysis performed: *accuracy*, *accessibility*, *usefulness*, *relevance*, and *security*.

3. Research Methodology

This section focuses on explaining the research methods applied in this research study. Initially, the rationale for the use of IHC as a research method is discussed. Subsequently, based on the hierarchical framework for organizing DQDs proposed by [13], four research questions are formulated.

3.1. Description of IHC

The IHC could be loosely described as the analysis and interpretation of texts and literature. The IHC consists of searching, sorting, selecting, acquiring, reading, identifying, and refining ideas within existing literature. IHC is reported to be applicable to research studies which are emergent in nature [33]. This property is reflected in the current work in this paper since the fields of Big Data and health informatics can be considered 'emerging' areas. As the discovery of DQDs most important to Big Data within the health industry is of a qualitative nature, potential research methods consist of the use of interviews of data quality managers and the integrative review of existing literature using the IHC. In this work, there are practical issues that deterred from the application of interviews as a research method, namely: the almost non-existence of data quality managers in Mauritius, which is where the research was carried out; the difficulty getting into contact with data quality managers who are external to Mauritius; and, most importantly, the fact that many data quality managers in the health industry have not yet adopted a proper framework in the context of Big Data.

On the other hand, IHC or similar integrative reviews of literature are research methods which have already been applied in the context of health and data [11] and in the context of Big Data quality [8]. This method is suitable for any domain of research which is emergent in nature and, as such, there are very few real-life or practical applications of the concepts discussed. This in-depth investigation of past literature allows for the formulation of theoretical frameworks which could be validated by further practical experiments once the emergent technology or domain area becomes more mainstream.

Therefore, despite the existence of very few academic publications related directly to data quality for Big Data in the health industry, an exploratory approach to data quality was undertaken in this work. Search activities were carried out on different research databases, with IEEExplore, ACM, health.gov, SCOPUS, Web of Science, and PubMed being the most well-known ones. The keywords and search operators used consisted of: data quality in big health datasets, data quality AND Big Data, data quality dimensions AND health datasets, information quality AND health datasets, very large datasets AND data quality, data streams, and data quality. Regarding the sorting phase, only articles published from the year 2006 to the present were selected. Other criteria for selection were: the popularity and hence acceptance of the article determined by the number of citations obtained, wherever possible; also, the interpretation of ideas or concepts put forward by authors and their relevance related to this current work. With regard to the latter, some work focusing on data quality with machine-related or sensor-based data was initially thought to be irrelevant for this research, but after analysis of some of those papers, corresponding ideas in terms of having similar types of data to the healthcare industry were discovered. Hence, those papers were included for further decoding and analysis.

3.2. Conceptual Reasoning

The main steps of the research carried out are depicted in Figure 1 below:

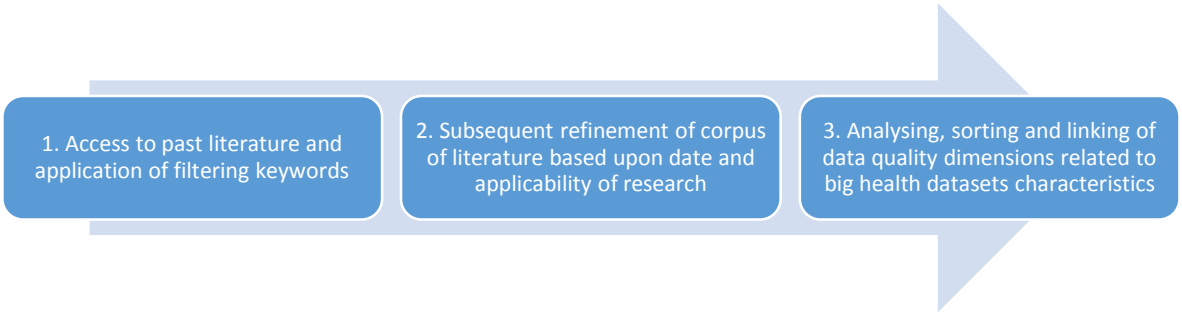

Figure 1. Main steps of inner hermeneutic cycle (IHC) applied.

The first two steps in Figure 1 are explained in the next section below. The analysis and linking of DQDs related to big health dataset characteristics were aimed at developing new knowledge. Part of the research methodology used in this work was adopted from previous work by Wang and Strong [13], i.e., mapping which category of DQD could be transposed to Big Data within the health industry context according to previous literature. The aim of Wang and Strong was to develop a hierarchical framework for organizing DQDs. They used clustering methods to associate a set of 15 dimensions into four categories according to the opinion of the participants. This clustering led to the labeling of the following categories, as discussed in Table 3:

Table 3. Data quality dimension (DQD) categories [13].

Category	Description and Main DQ Dimensions
Intrinsic	Is explained by data having quality in their own right. (accuracy, objectivity, believability, and reputation)
Contextual	Highlights the idea that data quality is a factor of the task at hand (value-added, relevancy, timeliness, completeness, and appropriate amount of data)
Representational	Includes aspects linked with the format and meaning of data (interpretability, ease of understanding, representational consistency, and concise representation)
Accessibility	Emphasizes the role of getting access to data (accessibility and access security)

Four important research questions, given below, were investigated.

RQ1: Should the Intrinsic category of DQDs still be applicable in the health industry context?

RQ2: How does the breadth of health data use cases impact the Contextual category of DQDs?

RQ3: Is the Representational category of DQDs negatively affected by the variety characteristic of Big Data and the fact that the quality of data would depend upon the aims of data analysis?

RQ4: Does the fact that health datasets are very often publicly available and voluminous data slows down access and retrieval of data negatively impact the Accessibility category of DQDs?

4. Work Undertaken

The initial search using the criteria discussed in Section 3 resulted in thousands of hits. With the SCOPUS database only, there were 2063 matching returns for the query "data quality and Big Data". However, subsequent manual analysis of abstracts of most of the matching articles and research studies resulted in less than 15 papers discussing data quality dimensions. This small amount confirmed the emerging nature of the area but also presented a practical issue in terms of having too few related research studies to perform IHC. Therefore, other search methods had to be devised to access relevant existing work. The term "health data" was appended to the original search criteria to search through journals and online resources focusing on health informatics.

Manual searching of some journals, such as *Data Science Central* and *Journal of Data and Information Quality (JDIQ)*, revealed that there were some potentially applicable research articles which were not

being highlighted by the search criteria mentioned above. For example, with JDIQ, out of around 160 matches with the broad key terms of "data quality", around 15 of them could be linked to either DQDs or data quality in health datasets or data quality in big datasets, but none of these 15 were matched when specifying the search criteria. This could be explained by different terms used as part of the titles of journal papers and the complexity of those titles, such as "Challenges in data quality: the influence of data quality assessments on data availability and completeness in a voluntary medical male circumcision programme in Zimbabwe". Some research work written before 2006 was also discovered to be relevant for the current research, and, therefore, it was included for the IHC. Even if it was out of the initial date range for the search of relevant literature, the current authors believed that the content from these research papers was still relevant and adequate and, therefore, should be included.

The source of some past literature also varied in terms of authority; not many sources originated from refereed journals or reviewed conference publications; therefore, articles and resources from credible health and data quality websites were consulted. For example, the Centre for Disease Control and Prevention (CDC) website generated over 100 matches just for the search criterion of 'data quality dimensions'. No additional results were returned using most of the other search terms mentioned at the beginning of this section. With the search term 'data quality in big health datasets', there were around 24 results, which were attributed greater weight for subsequent analysis. Results from some search terms from the CDC website were not research papers but reports and manuals, such as 'National Health and Nutrition Survey Anthropometry procedures manual'. These types of documentation were either not considered or considered with a low weight. Some of the results were pages which contained further links only to abstracts of papers but without the full paper details. In some cases, the abstract contained enough information relative to data quality dimensions and, therefore, the authors had to look for the complete paper. However, as CDC is a global center with high recognition and authority, the dimensions mentioned in the different reports were taken into account even if a lesser weight was given to those reports, compared to journals or conference papers. The final number of papers included for further reviewing amounted to 41.

The 42 papers retained for the integrative review were analyzed to understand and assess the importance of the DQDs discussed by the authors. An initial organization of the ideas is summarized in Table 4 below. A weight (L:Low, M:Medium, H:High) was assigned to each of the 41 papers. The weight was attributed taking into consideration the degree of alignment between the ideas presented within a paper and the context of the evaluation of DQDs for big datasets in the healthcare industry. This method of data evaluation was inspired by integrative review discussions in nursing, where reports were coded on a 2-point scale [34].

Low weights were assigned whenever the research lacked both the Big Data and health industry context, but some association could be made between the importance of DQDs cited and Big Data within health industry according to the authors' judgment. Medium weights were assigned when either a Big Data or health industry context was present. High weights were assigned when the two concepts (Big Data and health industry) were present. The judgment of the authors was again used to resolve semantic differences in jargon used by different authors to express specific data quality dimension.

Each weight was assigned a numerical value as follows: L: 1, M: 2, and H: 3. The total of weighted counts per DQD was computed to ascertain which DQDs were the most important following the application of the inner hermeneutic cycle. A staggering unique count of 43 distinct DQDs was noted following the IHC analysis. This confirmed the impression of a lack of a universal data quality framework and the possible fact that different authors might be using different jargon to express the same idea. DQDs cited only once with a low weight were discarded from further analysis, and the remaining number of DQDs became 38. The total count results of the dimensions listed tended to confirm previous general research on DQDs, with 'accuracy' being one of the most cited.

Table 4. Integrative review with details of weights.

Research Article USED	DQ Dimensions	Weight
(Panahy et al., 2013) [2]	Accuracy, Completeness, Consistency, Timeliness.	L
(Khan et al., 2012) [35]	Consistency, Completeness, Accuracy	M
(Jones et al., 2017) [36]	Accuracy	L
(Amoakoh-Coleman et al., 2015) [37]	Accuracy, Completeness	L
(Jacke et al., 2012) [38]	Accuracy, Completeness	M
(Langley et al., 2006) [29]	Accuracy	L
(Serhani et al., 2016) [39]	Accuracy, Completeness, Consistency, Timeliness	H
(Batini et al., 2006) [40]	Accuracy, Completeness, Accessibility, Trust, Readability, Consistency	M
(Xiao et al., 2017) [41]	Availability, Completeness	L
(O'Reilly et al., 2016) [30]	Accuracy, Capture, Completeness	M
(Giarrizzo-Wilson et al., 2011) [42]	Accuracy, Completeness, Trust, Legibility	L
(Giarrizzo-Wilson et al., 2011) [43]	Accuracy, Completeness	L
(Varshney et al., 2015) [31]	Accuracy, Privacy and Security, Heterogeneity, Provenance and Trust, Availability, Completeness	M
(Li et al., 2007) [44]	Completeness, Reliability, Correctness, Consistency, 'minimality'	L
(Sidi et al., 2012) [45]	Accuracy, Completeness, Consistency	L
(Weber et al., 2015) [46]	Correctness, Provenance, Currency, Plausibility	M
(Leon et al., 2016) [47]	Accuracy, Completeness, Consistency, Currency, Reliability, Uniqueness	H
(Pinto, 2006) [48]	Relevance, Consistency, Accuracy, Currency, Comprehensiveness, Format	L
(Arts et al., 2002) [49]	Accuracy, Completeness, Clarity, Format	M
(Weiskopf and Weng, 2012) [11]	Completeness, Correctness, Plausibility Concordance, Currency	M
(Geisler et al., 2011) [27]	Accuracy, Completeness, Consistency, Timeliness, Confidence, Data Volume	M
(Todoran et al., 2015) [50]	Accuracy, Completeness, Currency, Reliability	H
(Nystrom et al., 2013) [51]	Accuracy, Precision	L
(Huang et al., 2012) [10]	Accuracy, Usefulness, Accessibility, Relevance, Security	H
(Aday and Cynamon, 2010) [52]	Accuracy, Reliability, Consistency	H
(Cure, 2012) [53]	Accuracy, Completeness	M
(Vattulainen, 2015) [54]	Completeness, Redundancy, Accuracy, Representativeness, Consistency	L
(Byrd and Byrd, 2013) [55]	Accuracy, Completeness, Timeliness	M
(Lin et al., 2016) [56]	Completeness, Consistency, Coincidence	M
(Gibson, 1997) [57]	Accuracy, Completeness, Precision, Verifiability, Validity, Plausibility	M
(Vetro et al., 2016) [58]	Accuracy, Completeness, Understandability, Traceability, Compliance	M
(Craswell et al., 2016) [59]	Accuracy, Consistency, Clarity	M
(Lee and Haider, 2013) [60]	Believability, Security, Accuracy, Timeliness	L
(Lima et al., 2009) [61]	Reliability, Validity, Coverage, Accuracy, Completeness	M
(CDC, 2009) [62]	Consistency, Accuracy, Plausibility	M
(Cai and Zhu, 2015) [9]	Availability, Usability, Reliability, Relevance, Presentation Quality	M
(Blake and Mangiameli, 2011) [63]	Accuracy, Completeness, Consistency, Timeliness	L
(Salati et al., 2016) [64]	Completeness, Reliability	L
(Rhodegero, 2014) [65]	Completeness, Accuracy	M
(Kahn et al., 2012) [35]	Format, Availability, Timeliness, Consistency	L
(White, 2014) [66]	Consistency, Conciseness, Completeness, Expandability, Sensitivity.	M

5. Discussions

The results of the total of weighted counts per DQD are displayed in Table 5, sorted in descending order:

Table 5. Counts of different DQDs.

DQD	Total Weighted Count	DQD	Total Weighted Count
Accuracy	58	Readability	2
Completeness	52	Capture	2
Consistency	30	Privacy	2
Reliability	15	Heterogeneity	2
Timeliness	11	Provenance	2
Currency	8	Comprehensiveness	2
Availability	6	Concordance	2
Accessibility	5	Confidence	2
Trust	5	Data volume	2
Security	5	Coincidence	2
Correctness	5	Verifiability	2
Plausibility	4	Understandable	2
Relevance	4	Traceability	2
Clarity	4	Compliance	2
Validity	4	Coverage	2
Uniqueness	3	Usability	2
Format	3	Presentation quality	2
Precision	3	Expandability	2
Usefulness	3	Sensitivity	2

As seen in Table 5, it is clear that DQDs such as *accuracy* and *completeness* were the most cited, with 58 and 52 total counts, respectively. The authors posited that the dimensions with a total count of more than 10 are the ones which are most important in the context of Big Data within the health industry. Therefore, the most important DQDs are *accuracy*, *completeness*, *consistency*, *reliability*, *and timeliness*. This list of dimensions closely matches the 'information' section of the data quality framework proposed in the context of cloud-based health information systems [32]. Four out of five of the proposed dimensions are included in the 'information' section, with only *reliability* being excluded. In general, taking into account the discussions from the literature review, it is not surprising to see accuracy as the most important DQD. Accuracy is very often cited in most DQD research, but very often, the meaning of accuracy varies.

The results were analyzed within the framework of [13], discussed earlier in Section 3.2. Table 6 below shows the summation of individual DQD counts when grouped within the four categories advocated by the hierarchical framework for organizing DQDs developed by Wang and Strong [13].

Table 6. DQ category dimensions with count aggregates.

Category	Individual Dimensions	Count
Intrinsic	Accuracy, Trust, Plausibility, Precision, Compliance, Traceability, Verifiability, Provenance, Confidence, Concordance, Correctness	87
Contextual	Completeness, Timeliness, Currency, Reliability, Availability, Uniqueness, Relevance, Validity, Expandability, Sensitivity, Coverage, Data volume, Comprehensiveness, Heterogeneity	108
Representational	Consistency, Format, Usefulness, Readability, Capture, Coincidence, Understandable, Usability, Presentation quality	48
Accessibility	Accessibility, Security, Privacy, Compliance	14

Hence, the following might be concluded for each of the research questions:

RQ1: Should the 'Intrinsic' category of DQDs still be applicable in the health industry context?

The 'Intrinsic' category ranks second, as per Table 5. This suggests that some DQDs, such as *accuracy* and *trust* in data, are applicable in all situations in which data could be used, including that of Big Data for the health industry. However, at the same time, the fact that the 'Intrinsic' DQD category is not the most cited category gives some credit to previous researchers who state that data quality might not be impactful in the context of Big Data [26]. Finally, it also points to the fact that data quality work in the health industry context should not be considered intrinsically, but the importance of DQDs vary according to software applications accessing data and users working with data.

RQ2: How does the breadth of health data use cases impact the Contextual category of DQDs?

The Contextual category dimension carries the highest importance. The results of the IHC imply that DQDs in the Contextual category have a higher collective importance for Big Data in the health industry. This might be explained by the fact that there is such a huge variety of categories of data (patient-related, genomic, trauma-based, and others) and many different end-consumers of data (doctors, insurance companies, pharmaceutical groups, and others) such that each specific use of data might uphold different perspectives of quality to suit the "fitness for use" definition of data quality. Hence, the conclusion is that the context is extremely important for data quality applications in the specific domains of Big Data within the health industry.

RQ3: Is the Representational category of DQDs negatively affected by the variety characteristic and the fact that the quality of data would depend upon the aims of data analysis?

The importance of the Representational DQ category is quite low relative to the two previous categories with a total count of only 48. The authors were of the opinion that this category could be relatively unknown due to the *variety* characteristic of Big Data. However, most of the research data involved in the IHC conducted in this work concerned mostly text-based data, including numbers, and the results were analyzed and used by well-trained personnel. This could warrant future work concerning data quality with other kinds of data (images, charts, and videos) used in the health industry. On the other hand, even if some authors, such as [23], point toward the final goals of data analysis and how data quality is therefore important, most other research studies as part of the IHC do not give enough indication pertaining to the rationale behind the data analysis. Therefore, there is the inherent assumption in most research studies that data would be used in one or very few use cases; as discussed in the earlier sections of this paper, this assumption might not hold true, especially in the Big Data context.

RQ4: Does the fact that health datasets are very often publicly available and voluminous data slows down access and retrieval of data negatively impact the Accessibility category of DQDs?

Finally, the Accessibility category ranks unsurprisingly last in the findings. Many datasets are publicly available for analysis and research; hence, previous authors have undertaken their work within a context in which accessibility was easy. Thus, the volume aspect could be deduced not to have affected the accessibility to data, but care should be taken to probe the investigation of the relationship between *volume* and *accessibility* for Big Data applications. Furthermore, for private and industry-based applications of Big Data within the health industry, it could be argued that DQDs such as Security, Privacy, and Compliance would have a higher impact. Thus, future Big Data governance frameworks specifically for the health industry should further explore this Accessibility category.

6. Conclusions and Implications

This research was set in a multidisciplinary context involving three main fields: data quality (as an integral part of data governance), Big Data, and health informatics. Although data quality has been a well-researched field for the last two decades, there is still a lack of precise nomenclature when it comes

to data quality dimensions. Big Data and health informatics are two emerging fields with very little past research focusing on the data quality perspective. The use of DQDs as part of data governance initiatives is arguably very important, as it ensures that data users can extract the maximum value from data use. The aim of the research discussed in this paper was to investigate methodically which DQDs could be more important within the context of Big Data in the health industry. Initial reviews of literature suggested that previous research for this specific field is extremely limited. The inner hermeneutic cycle (IHC) was adopted as a research method due to the emerging degree of novelty of Big Data and health informatics, but also due to the impracticality of implementing other research methods because of organizational, geographical, legal, and ethical constraints.

The results confirm the popularity of some well-discussed DQDs, such as *Accuracy*, *Completeness*, *and Consistency.* Based on this research, the DQDs of *reliability and timeliness* were added to the three aforementioned DQDs to make up a list of the five most adequate DQDs for the context of Big Data within the health industry. When mapping the results of the IHC to the hierarchical DQ category framework of Wong and Strong [13], the Contextual category of DQDs was considered to be the most important. This could be easily explained by the broadness of the three domains (i.e., Big Data, data quality, and health informatics) involved, where there could be thousands of unique applications of Big Data for the health industry. Thus, for each application, the probability of selecting different DQ dimensions increases.

This current research study provides a hierarchy of the most important DQDs in the specific context of Big Data within the healthcare industry. It is one of the first studies within this area to use a systematic method to achieve this. This research study can be further validated by the use of the same research methods but applied in different contexts. Taking into account the very contextual nature of DQ, it would be expected that different contexts would produce a different list of the most important DQDs.

The comparatively small volume of literature, and, more specifically, papers denoting a high weight, is one of the main limitations of this research study. Other research methods which could be applicable using the same corpus of literature are being planned, with the main aim of reducing the amount of author bias introduced when considering the weights allocated to papers.

This research study will serve as a foundation for further research in the context of Big Data in the health industry to be conducted by the authors. Using the DQDs as features for machine learning algorithms, future work will distinguish quality data from non-quality data from very large streams of health datasets. Also, the application of more enhanced statistical methods, such as least semantic analysis (LSA), for determining the most important DQD dimensions will be used in further research. Further, empirical research methods will be applied with a wider choice of data variety in order to probe deeper into the Representational and Accessibility data quality categories (derived from Strong and Wong's framework described above) to confirm the findings of this current research. Finally, the five main DQ dimensions determined would serve as a basis to help determine which machine learning algorithms could be more efficient in classifying incorrect data. Following this, an optimal data repair algorithm which will not be computationally expensive will be developed.

Author Contributions: S.J. acted as principal researcher and main manuscript author. C.G. was principal advisor and reviewer and also contributed in manuscript corrections. P.D. and D.W. acted as advisor and reviewer and contributed in manuscript corrections.

Funding: This research received no external funding.

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

References

1. O'Driscoll, A.; Daugelaite, J.; Sleator, R.D. 'Big data', Hadoop and cloud computing in genomics. *J. Biomed. Inform.* **2013**, *46*, 774–781. [CrossRef] [PubMed]
2. Panahy, P.; Sidi, F.; Affendey, L.; Jabar, M.; Ibrahim, H.; Mustapha, A. Discovering dependencies among data quality dimensions: A validation of instrument. *J. Appl. Sci.* **2013**, *13*, 95–105. [CrossRef]
3. Raghupathi, V.; Raghupathi, W. Big data analytics in healthcare: Promise and challenges. *Health Inf. Sci. Syst.* **2014**, *2*, 3. [CrossRef] [PubMed]
4. Handler, D.J. Small Data-Thinking Kills Big Data-Aspirations. Available online: http://www.wired.com/insights/2013/01/small-data-thinking-kills-big-data-aspirations/ (accessed on 27 April 2017).
5. Sackett, D.L.; Rosenberg, W.M.; Gray, J.A.; Haynes, R.B.; Richardson, W.S. Evidence based medicine: What it is and what it isn't. *BMJ* **1996**, *312*, 71–72. [CrossRef] [PubMed]
6. Yesha, Y.; Janeja, V.; Rishe, N.; Yesha, Y. Personalized Decision Support System to Enhance Evidence Based Medicine through Big Data Analytics. In Proceedings of the 2014 IEEE International Conference on Healthcare Informatics (ICHI), Verona, Italy, 15–17 September 2014.
7. Zolfaghar, K.; Meadem, N.; Teredesai, A.; Roy, S.B.; Chin, S.C.; Muckian, B. Big data solutions for predicting risk-of-readmission for congestive heart failure patients. In Proceedings of the 2013 IEEE International Conference on Big Data, Santa Clara, CA, USA, 6–9 October 2013; pp. 64–79.
8. Batini, C.; Rula, A.; Scannapieco, M.; Viscusi, G. From data quality to big data quality. *J. Database Manag.* **2015**, *1*, 60–82. [CrossRef]
9. Cai, L.; Zhu, Y. The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. *Data Sci. J.* **2015**, *14*, 2. [CrossRef]
10. Huang, H.; Stvilia, B.; Bass, H. Prioritization of Data Quality Dimensions and Skills Requirements in Genome Annotation Work. *J. Am. Soc. Inf. Sci. Technol.* **2012**, *63*, 195–207. [CrossRef]
11. Weiskopf, N.G.; Weng, C. Methods and dimensions of electronic health record data quality assessment: Enabling reuse for clinical research. *J. Am. Med. Inf. Assoc.* **2013**, *20*, 144–151. [CrossRef] [PubMed]
12. Juddoo, S. Overview of Big Data quality challenges. In Proceedings of the IEEE ICCCS Conference, Pointe au Piments, Mauritius, 4–5 December 2015. [CrossRef]
13. Wang, R.; Strong, D. Beyond Accuracy: What Data Quality Means to Data Consumers. *J. Manag. Inf. Syst.* **1996**, *12*, 5–33. [CrossRef]
14. Russom, P. Data Governance Strategies. *Bus. Intell. J.* **2008**, *13*, 13–14.
15. Thomas, G. Data Governance Institute. Available online: www.DataGovernance.com (accessed on 8 July 2017).
16. Barakhan, B. Drive towards Data Governance. Available online: http://www.ewweb.com (accessed on 11 July 2016).
17. IBM. IBM Data Governance Council Maturity Model. Available online: www.ibm.com/software/data/information/trust-governance.html (accessed on 7 May 2016).
18. Khatri, V.; Brown, C. Designing data governance. *Commun. ACM* **2012**, *53*, 148–152. [CrossRef]
19. Malik, P. Governing Big Data: Principles and Practices. *IBM J. Res. Dev.* **2013**, *57*, 1. [CrossRef]
20. Pipino, L.; Yang, L.; Wang, R. Data Quality Assessment. *Commun. ACM* **2002**, *45*, 211–218. [CrossRef]
21. Islam, M.S. An Assessment for focusing the change of data quality (DQ) with timeliness in Information Manufacturing systems. In Proceedings of the SDIWC Second International Conference on Digital Enterprise and Information Systems (DEIS2013), Kuala Lumpur, Malaysia, 4–6 March 2013.
22. Cappiello, C.; Francalanci, C.; Pernici, B. Data quality assessment from the user's perspective. In Proceedings of the International Workshop on Information Quality in Information Systems, Paris, France, 18 June 2004.
23. Caballero, I.; Serrano, M.; Piattinni, M. A Data Quality in Use Model for Big Data. In *ER 2014: Advances in Conceptual Modeling*; Springer: Cham, Switzerland, 2014; pp. 65–74.
24. Dong, X.L.; Srivastava, D. *Big Data Integration*; Morgan and Claypool Publishers: San Rafael, CA, USA, 2015. [CrossRef]
25. Saha, B.; Srivastava, D. Data Quality: The other face of Big Data. In Proceedings of the 2014 IEEE 30th International Conference on Data Engineering (ICDE), Chicago, IL, USA, 31 March–4 April 2014.
26. Soares, S. *Big Data Quality, Big Data Governance: An Emerging Imperative*; MC Press: Boise, ID, USA; pp. 101–112.

27. Geisler, S.; Weber, S.; Quix, C. Ontology-Based Data Quality Framework for Data Streams. In Proceedings of the 16th International Conference on Information Quality, Adelaide, Australia, 18–20 November 2011.
28. CIHI. Canadian Institute for Health Information. Available online: https://www.cihi.ca/en/data-and-standards/data-quality (accessed on 20 August 2018).
29. Langley, J.; Stephenson, S.; Thorpe, C.; Davie, G. Accuracy of injury coding under ICD-9 for New Zealand public hospital discharges. *Injury Prev.* **2006**, *12*, 58–61. [CrossRef] [PubMed]
30. O'Reilly, G.; Gabbe, B.; Moore, L.; Cameron, P. Classifying, measuring and improving the quality of data in trauma registries: A review of literature, Injury. *Int. J. Care Injured* **2016**, *47*, 559–567. [CrossRef] [PubMed]
31. Varshney, K.R.; Wei, D.; Ramamurthy, K.N.; Mojsilovic, A. Data Challenges in Disease Response: The 2014 Ebola Outbreak and beyond. *ACM J. Data Inf. Qual.* **2015**, *6*, 2–3. [CrossRef]
32. Almutiry, A.; Wills, G.; Alwabel, A.; Crowder, R.; Walters, R. Toward A Framework For Data Quality in Cloud-Based Health Information System. In Proceedings of the IEEE International Conference on Information Society, London, UK, 10–12 November 2014.
33. Richard, T. Writing Integrative Literature Reviews: Guidelines and Examples. *Hum. Resour. Dev. Rev.* **2005**, *4*, 356–367. [CrossRef]
34. Whittemore, R.; Knafl, K. The integrative review: Updated methodology. *J. Adv. Nurs.* **2005**, *5*, 546–553. [CrossRef] [PubMed]
35. Kahn, M.; Raebel, M.A.; Glanz, J.M.; Riedlinger, K.; Steiner, J. A Pragmatic Framework for Single-site and Multisite Data Quality Assessment in Electronic Health Record-based Clinical Research. *Med. Care* **2012**. [CrossRef] [PubMed]
36. Jones, K.; Zenk, S.; Tarlov, E.; Powell, L.; Matthews, S.; Horoi, I. A step-by-step approach to improve data quality when using commercial business lists to characterize retail food environments. *BMC Res. Notes* **2017**, *10*. [CrossRef] [PubMed]
37. Amoakoh-Coleman, M.; Kayode, G.A.; Brown-Davies, C.; Agyepong, I.A.; Grobbee, D.E.; Klipstein-Grobusch, K.; Ansah, E.K. Completeness and accuracy of data transfer of routine maternal health services data in thegreater Accra region. *BMC Res. Notes* **2015**. [CrossRef] [PubMed]
38. Jacke, C.O.; Kalder, M.; Wagner, U.; Albert, U. Valid comparisons and decisions based on clinical registers and population based cohort studies: Assessing the accuracy, completeness and epidemiological relevance of a breast cancer query database. *BMC Res. Notes* **2012**, *5*, 700. [CrossRef] [PubMed]
39. Serhani, M.A.; Kassabi, H.T.; Taleb, I.; Nujum, A. An Hybrid Approach to Quality Evaluation across Big Data Value Chain. In Proceedings of the 2016 IEEE International Congress on Big Data, San Francisco, CA, USA, 27 June–2 July 2016.
40. Batini, C.; Scanapieca, M. Data Quality. In *Data Centric Systems and Applications*; Springer: Berlin, Germany, 2006.
41. Xiao, Y.; Bochner, A.F.; Makunike, B.; Holec, M.; Xaba, S.; Tshimanga, M.; Chitimbire, V.; Barnhart, S.; Feldacker, C. Challenges in data quality: The influence of data quality assessments on data availability and completeness in a voluntary medical male circumcision programme in Zimbabwe. *BMJ Open* **2017**, *7*, e013562. [CrossRef] [PubMed]
42. Giarrizzo-Wilson, S.; Maxwell-Downing, D.; Bianco, J. Data quality and the electronic health record (EHR). *AORN J.* **2011**, *94*. [CrossRef]
43. Giarrizzo-Wilson, S.; Maxwell-Downing, D.; Bianco, J. Pre-charting patient care information. *AORN J.* **2011**, *94*. [CrossRef]
44. Li, X.; Shi, Y.; Li, J.; Zhang, P. Data Mining Consulting Improve Data Quality. *Data Sci. J.* **2007**, *6*, S658–S666. [CrossRef]
45. Sidi, F.; Jabar, M.; Ibrahim, H.; Mustapha, A. Data Quality: A survey of data quality dimensions. In Proceedings of the 2012 International Conference on Information Retrieval & Knowledge Management, Kuala Lumpur, Malaysia, 13–15 March 2012. [CrossRef]
46. Weber, J.; Price, M.; Davies, I. Data Quality by Contract—Towards an Architectural View for Data Quality in Health Information Systems. In Proceedings of the Conference on Artificial Intelligence in Medicine in Europe, Pavia, Italy, 17–20 June 2015; pp. 143–157. [CrossRef]
47. Leon, A.; Reyes, J.; Burriel, V.; Valverde, F. Data Quality Problems when Integrating Genomic Information. In Proceedings of the ER 2016 Workshops, Gifu, Japan, 14–17 November 2016; pp. 173–182. [CrossRef]

48. Pinto, M. Data representation factors and dimensions from the quality function deployment (QFD) perspective. *J. Inf. Sci.* **2006**. [CrossRef]
49. Arts, D.G.T.; De Keizer, N.F.; Scheffer, G. Defining and Improving Data Quality in Medical Registries. *Am Med. Inform. Assoc.* **2002**, *9*, 600–611. [CrossRef]
50. Todoran, I.-G.; Lecornu, L.; Kenchaf, A.; Le Caillac, J.M. A Methodology to Evaluate Important Dimensions of Information quality in systems. *J. Data Inf. Qual.* **2015**, *6*, 11. [CrossRef]
51. Nystrom, M.; Andersson, R.; Holmqvist, K.; Weijer, J. The influence of calibration method and eye physiology on eyetracking data quality. *Behav. Res.* **2013**, *45*, 272–288. [CrossRef] [PubMed]
52. Aday, L.A.; Cynamon, M. Health Survey Research Methods. In Proceedings of the 9th Conference on Health Survey Research Methods, Centers for Disease Control and Prevention, National Center for Health Statistics, Hyattsville, MD, USA, May 2010.
53. Cure, O. Improving the data quality of drug databases using conditional dependencies and ontologies. *ACM J. Data Inform. Qual.* **2012**, *4*, 3. [CrossRef]
54. Vattulainen, M. Improving the Predictive Power of Business Performance measurement systems by Constructed Data Quality Features? Five cases. In *Advances in Data Mining: Applications and Theoretical Aspects*; Springer: Cham, Switzerland, 2015; pp. 3–16. ISBN 978-3-319-20910-4.
55. Byrd, W.L.; Byrd, T.A. Contrasting the Dimensions of Information Quality in their Effects on Healthcare Quality in Hospitals. In Proceedings of the 46th Hawaii International Conference on System Sciences, Wailea, Maui, HI, USA, 7–10 January 2013.
56. Lin, Y.; Wang, H.; Zhang, S.; Li, J.; Gao, H. Efficient quality-driven source selection from massive data sources. *J. Syst. Softw.* **2016**, *118*, 221–233. [CrossRef]
57. Gibson, N. *Measuring the Quality of Patient Data with Particular Reference to Data Accuracy*; University of Keele: Keele, UK, 1997.
58. Vetro, A.; Canova, L.; Torchiano, M.; Iemma, R. Open data quality measurement framework: Definition and application to Open Government Data. *Gov. Inf. Q.* **2016**, *33*, 325–337. [CrossRef]
59. Craswell, A.; Moxham, L.; Broadbent, M. Does use of computer technology for perinatal data collection influence data quality? *Health Inform. J.* **2016**, *22*, 293–303. [CrossRef] [PubMed]
60. Lee, H.S.; Haider, A. Identifying Relationships of Information Quality Dimensions. In Proceedings of the 2013 Technology Management for Emerging Technologies (PICMET'13), San Jose, CA, USA, 28 July–1 August 2013.
61. Lima, C.R.; Schramm, J.M.; Coeli, C.M.; da Silva, M.E. Review of data quality dimensions and applied methods in the evaluation of health information systems. *Cad. Saúde Pública* **2009**, *25*, 2095–2109. [CrossRef] [PubMed]
62. CDC. Joint Meeting of International Collaborative Effort on Injury Statistics and the Global Burden of Disease-Injury Expert Group. Available online: http://www.cdc.gov/nchs/injury/ice/boston2009/boston2009_proceedings.htm#proceeding_14 (accessed on 13 January 2017).
63. Blake, R.; Mangiameli, P. The Effects and Interactions of Data Quality and Problem Complexity on Classification. *ACM J. Data Inf. Qual.* **2011**, *2*. [CrossRef]
64. Salati, M.; Falcoz, P.E.; Decaluwe, H.; Rocco, G.; Van Raemdonck, D.; Varela, G.; Brunelli, A.; ESTS Database Committee. The European thoracic data quality project: An Aggregate Data Quality score to measure the quality of international multi-institutional databases. *Eur. J. Cardiothorac. Surg.* **2016**, *49*, 1470–1475. [CrossRef] [PubMed]
65. Rhodegero, J. The use of big data in manual physiotherapy. *Man. Ther.* **2014**, *19*, 509–510. [CrossRef] [PubMed]
66. White, P. A Pilot Ontology for a Large, Diverse Set of National Health Service Healthcare Quality Indicators. Unpublished Ph.D. Thesis, City University London, London, UK, 2014.

© 2018 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).