International Journal of Engineering Research and Development
e-ISSN: 2278-067X, p-ISSN: 2278-800X, www.ijerd.com
Volume 13, Issue 9 (September 2017), PP.46-54

Effective Data Quality Management In Health Sector

*K.Parish Venkata Kumar [1],Dr. B.D.C.N Prasad [2]
[1]*Research Scholar, Dept of Computer Science and Engineering, Rayalaseema University, Kurnool.*
[2]*Dept of Computer Science and Engineering PSCMR College of Engineering & Tech, Vijayawada.*
*Corresponding Author: *K.Parish Venkata Kumar*

ABSTRACT:- Data quality analysis leavings a difficult issue on several sphere (e.g. geographic, software, databases, etc.). This is especially the case on e-Health control applications for continuing of data aspect to ensure correct decision making is very crucial. Patients monitoring assign to a continued observation of patient's quality (physiological and physical) traditionally achieve by one or several body sensors. In fact, compelling actions and compromise are based on data coming from such sensors (e.g. remote diagnosis, consultations, hospitalization…). Contribute high data quality helps to assurance a correct processing and analysis of information, as well as the applicable interference of medical services. In this paper, we explore the assumptions and issues of data quality in this appropriate domain providing primary research indication and motivation about this conditional. We underline the obligation of the analysis of data character on e-Health applications, exclusively regarding remote.

Keywords:- Data quality ,quality management ,E-health ,decision making, data analysis.

I. INTRODUCTION

According to the WHO (World Health Organization) in 2020 most of the diseases worldwide will be due to chronic genetics as diabetes, hypertension or cardiovascular diseases. Thus aggregate the problems of over weight and intensify the enterprise monitoring (i.e. actimetry). The progression of such physiology requires numerous and overpriced cares. Home care combine to a remote medical audit and assistance becomes necessary. Nowadays, the advancement of ICT (Information and publicity Technology) vigorously helps to provide better character of healthcare. For example, the use of high-technology body sensors (i.e. pulse, body temperature, ECG…), wired and wireless communications automation, real-time data convert, interactive consolidate, etc. This advance has been a encouragement for new healthcare programs and access (i.e. Medic4you, Health Guide, Medmobile…) which analysis to better assist patients with chronic or actual diseases. Such programs allow better quality and convenience of healthcare systems and develop the message exchange between medical specialists. However, the administration of data in this kind of organization is becoming progressive complex. Periodically, decision makers (medical experts or professionals, medical services…) are oppose to inaccurate, inadequate or excessive intelligence. As a result, more and more questions about data quality, security and confidentiality in this domain arise. Particularly, assure the data quality in healthcare domain leavings an important argument. If data quality is avoid, confident data may have greatly negative impact on the attainment of the application and on the arrangement making. In this analysis work, we claim that data element in e-Health control applications cannot be ignored and neither prescribed to basic data quality access. We believe that a better considerate of the meaning of data quality issues develops also the quality of controlling and thus better will be the patient conclusion. Several appearance of data quality search over e-Health monitoring operation are illustrated in this paper by a scheme from a current analysis project – STM3: A solution for the medical compensations and observe in a mobile context - grouping technical and academic research teams, as well as users and corporation from electronics ,publicity , and computer science territory. The project is financed by the French cluster SCS (Secure Communication Solutions).

II. MOTIVATION

In the last few years, professional improvement opens new capabilities to healthcare and medicine practice, but carriers some derive risks and leave decision makers with various unanswered questions about quality, guarantee and other important matters. Some surveys and access have showed the consequence of data quality of end-users, especially in healthcare domain. E-Health control applications have some particularities regarding the importance on data quality. On the one hand, fortunate healthcare delivery and plan strongly rely on data (e.g. sensed data, diagnosis, administration information); the higher quality of the data, the better will be the patient benefit. On the other hand, these operations are also particularly exposed to a dependent environment (i.e. patients' mobility, connection technologies performance, information heterogeneity…) that has an important impact on intelligence administration and application achievement. Motivated by these considerations, we study the related data quality issues over the precision of e-Health monitoring applications.

III. DATA QUALITY LITERATURE

'Data quality' and then arrange a fundamental understanding of the shock of poor quality data. Finally, the section consider existing models of the communication between data preservation effort and costs inflicted by poor quality.Data quality is often defined as 'fitness for use', i.e. an calculation of to which extent some data serve the aspiration of the user (e.g. Lederman et al., 2003; Tayi& Ballou, 1998; Watts &Shankaranarayanan, 2009). Another way to accept the concept of data quality is by dividing it into subcategories and capacity. An often cited definition is implementing by Ballou and Pazer (1985), who divide data quality into four capacity: accuracy, timeliness, completeness, and flexibility. They argue that the certainty dimension is the easiest to checkout as it is merely a matter of analysing the difference among the correct value and the certain value used. They also argue that the evaluation of opportunity can be implement in a similar un problematic manner. As for the evaluation of the integrity of some data, this can also be done approximately straight forward, as long as the target is on whether the data are entire or not in contrast to defining the level of completeness, e.g. the interest of data completeness. On the other hand, an evaluation of flexibility is a little more complex, since this desire two or more portrayal schemes in order to be able to make a correlation.

Another data quality allocation is provided by Wand and Wang (1996). They limit their focus to inherent data qualities, of which they define four intrinsic amplitudes: completeness, unambiguousness, expressiveness and correctness. Wand and Wang (1996) take as their basis a paper, which appearance a review of cited data quality amplitude, i.e. the comprehensive information review of Wang et al. (1995). Wang et al. (1995) summarize the most often cited data condition dimensions. Wang and Strong (1996) come up with dataquality arrangements which divides data quality into four categories: intrinsic, circumstantial, representational, and convenience. For each category they define a set of amplitude, 18 in all. The definition by Wang and Strong (1996) is consider by Haug et al. (2009) who argue that 'descriptive data quality' can be perceived as a form of 'convenience data quality' instead of a division of its own. Thus, Hauget al. (2009) defines three data quality division: intrinsic, openness and usefulness. Levitin and Redman (1998) provide another attitude by arguing that since development to produce data have many comparison to processes that produce physical products, data producing action could be viewed as generating data products for data purchaser. With a basis in this view of data as resources, Levitin and Redman discuss how thirteen basic properties of departmental budget may be adapted into properties for data.

Impacts of poor quality data

The development of intelligence technology during the last decade has enabled management to collect and store excessive amounts of data. However, as the data volumes development, so does the complication of managing them. Since larger and more complex data resources are being collected and educated in organizations today, this equipment that the risk of poor data quality increases (Watts &Shankaranarayanan, 2009). Another often specified data related problem is that association often manage data at a local level (e.g. department or location). This implies the creation of ‘material silos’ in which data are constantly stored managed and handled (Lee et al., 2006; Smith, 2008; Vayghan et al., 2007). In this vein, Lee et al. (2006) argue that data silos imply that many association face a multitude of difference in data definitions, data formats and data values, which makes it almost hopeless to find out and use key data. From a explanation perspective, ERP systems have been advocate as a panacea for dealing with the lack of data mixture by replacing partially coordinated legacy arrangements (Davenport, 1998; Knolmayer&Röthlin, 2006). However, it has been proposed that data problems may get enhance when using ERP systems since the ERP modules are intricately linked to each other, which is the reason why poor quality data input in one width can affect the operating of other schedule negatively (Park &Kusiak, 2005).

Poor quality data can imply a aggregation of negative importance in a company. To start with, poor quality data that is not importance and corrected can have significantly negative economic and social impacts on a grouping (Ballou et al., 2004; Wang & Strong, 1996). The indication of poor quality data carry adverse effects to business users over: less customer satisfaction,expanded running costs, inefficient decision-making development, lower performance and lowered representative job satisfaction (Kahn et al., 2003; Leo et al., 2002; Redman, 1998). Poor data quality also expansion useful costs since time and other effects are spent detecting and revise errors. Since data are design and used in all daily operations, data are demanding inputs to almost all agreement and data implicitly define common terms in an enterprise, data establish a significant grantor to organizational culture. Thus, poor data quality can have adverse effects on the organizational experience (Levitin & Redman, 1998; Ryu et al., 2006). Poor data aspect also means that it develop into difficult to build trust in the company data, which may imply a lack of user acceptance of any action based on such data.

Hen concentrate on clarifying the effects of poor aspect data, it is clear that many association experience significant costs as a consequence of poor quality data, although the exact expansion of such costs is crucial to estimate. According to Redman (1998), consideration to produce estimates of the total cost of poor data quality have confirm difficult to perform. Additional data quality research has not yet progressive to the

point of having accepted measurement methods for any of this concern. On the other hand, Redman (1998) application that many case studies aspect accuracy allowance, but he does not provide quotation or mentions if these are intellectual studies. According to Redman (1998), measured at the field level, the announced error rates are in the interval of 0.5–30%. Furthermore, Redman (1998) claims that at least three medication studies have yielded estimates in the 8-12% of revenue range, but informally 40-60% of the expense of the assistance organization may be dominate as a result of poor data. Much demonstrate that the economic effect of even small data defect can be very significant. HäkkinenandHilmola (2008) argue that insignificant data blunder (e.g. 1-5%) may not automatically represent a major problem in construction, but that such inaccuracies will have direct effects in terms of lost sales and operational interruption in the after-sales organizations.In contrast to the possible lack of large consideration of data quality in intellectual journal papers (Eppler &Helfert, 2004; Kim & Choi, 2003), many industry experts arrange such studies. These industry authority include Gartner Group, Price Waterhouse Coopers and The Data gather Institute, which claim to determine a crisis in data quality management and a hesitation among senior decision-makers to do enough about it (Marsh, 2005). Marsh (2005) compile the findings from such surveys into the subsequent bullet-points (quoted from: Marsh, 2005):

•"88 per cent of all data combination projects either fail completely or significantly over-run their budgets"
•"75 per cent of organisations have described costs stemming from dirty data"
•"33 per cent of organisations have delayed orabolish new IT systems because of poor data"
•"$611bn per year is lost in the US in poorly target mailings and staff aerial alone"
•"According to Gartner, bad data is the upward one cause of CRM system failure"
•"Less than 50 per cent of companies claim to be very confident in the quality of their data"
•"Business brilliance (BI) projects often fail due to dirty data, so it is compulsory that BI-based business choice are based on clean data".

IV. METHODOLOGY

The Data Quality Cycle

Figure 1: The Data Quality Cycle

Before you can fully understand every component of the Data Quality Cycle, you must first accept the Measure composing and how Data Quality is measured analysis of Data Quality.

Every management is unique, but there are a number of significant Data Quality allowances that are universal:

•**Completeness:** The degree to which all required circumstance of data are occupy.
•**Anotherness:** The extent to which all distinct values of a data component appear only once.
•**Validity:** The amplification of how a data value coordinate to its territory value set (i.e., a set of acceptablevalues or range of values).
•**Accuracy:** The degree of observance of a data element or a data set to an accurate source that is assume to becorrect or the degree the data accurately perform the truth about a real-world object.
•**Integrity:** The degree of consent to defined data relationship guideline (e.g., primary/foreign key referentialintegrity).
•**Timeliness:** The degree to which data is accessible when it is appropriate.
•**Consistency:** The degree to which a exclusive piece of data holds the same value across different data sets.

•**Representation:** The indicative of Data Quality that location the format, pattern, legibility, and efficiency of data for its calculated use.

In addition to significant Data Quality measures, approximate measures should also be considered. Some illustrations include.

•**Business amusement Measures:** The increase/decrease in business satisfaction based on analysis.
•**Collaboration/Improved capacity Measures:** Percent of times the Data administration Council detected and dispose of redundant intra- or inter-managerial projects/initiatives.
•**Business convenience/Risk Measures**: Business benefit achieve due to quality data or employment risk realized due to debatable data. Increase in competitive data due to data availability and Data Quality advance.
•**Compliance allotment:** Users with access to update/consequence the master data are confined to only those employees who have need and have been accepted as part of their job functions.

It is very important to inaugurate the measures of Data Quality most crucial to your organization. This is required to establish a control for the quality of your data and to monitor the growth of your DQM initiatives. The other foundational composing of the Data Quality Cycle appropriate to Discover, Profile, Establish Rules, Monitor, Report, Remediate, and continuously improve Data Quality are described in the next section.

Components of DQM
Once In Place, These Key Components Grant Robust, Recyclable And Highly Effective DQM Capabilities That Can Be Leveraged Across The Enterprise:
•Data Discovery: The process of finding, association, organizing and reporting metadata about your data (e.g., files/tables, record/row definitions, field/column definitions, keys).
•Data Profiling: The process of consider your data in detail, comparing the data to its metadata, considerate data statistics and reporting the allotment of quality for the data at a point in time.
•Data Quality Rules: Based on the business demand for each Data Quality measure, the business and technical rules that the data must observe to in order to be examined of high quality.
•Data Quality Monitoring: The ongoing check of Data Quality, based on the results of eliminate the Data Quality rules, and the connection of those results to defined error entrance, the creation and storage of Data Quality reservation and the generation of appropriate proclamation.
•Data Quality Reporting: The reporting, dashboards and yellow pages used to report and trend ongoing Data Quality allotments and to drill down into detailed Data Quality omission.
•Data Remediation: The ongoing alteration of Data Quality exceptions and concern as they are reported.
Each of these DQM components is characterize in greater detail in terms of roles and responsibilities, processes, automation and business benefits in the sections that follow.

Data Discovery

Roles and Responsibilities
Data discovery is frequently the responsibility of IT. However, tech-savvy business users/managers may also perform data revelation when user-friendly data discovery tools are available.

Processes

Data discovery should be an computerized process using a robust data discovery tool. The data domains and physical database servers and/or file systems in capacity must first be identified and read-only security acknowledgment to those database servers and/or file systems must be accomplish in order to execute the discovery processes. The discovery tool will gather all of the available metadata and store it in a discovery metadata archive where it can then be queried and analyzed. The metadata grab typically include database .Schema/file directory names, table/file names and definitions, column/field names and definitions, and any defined database or file relationships (e.g., primary/foreign key relationships).Technologies.

V. RESULTS

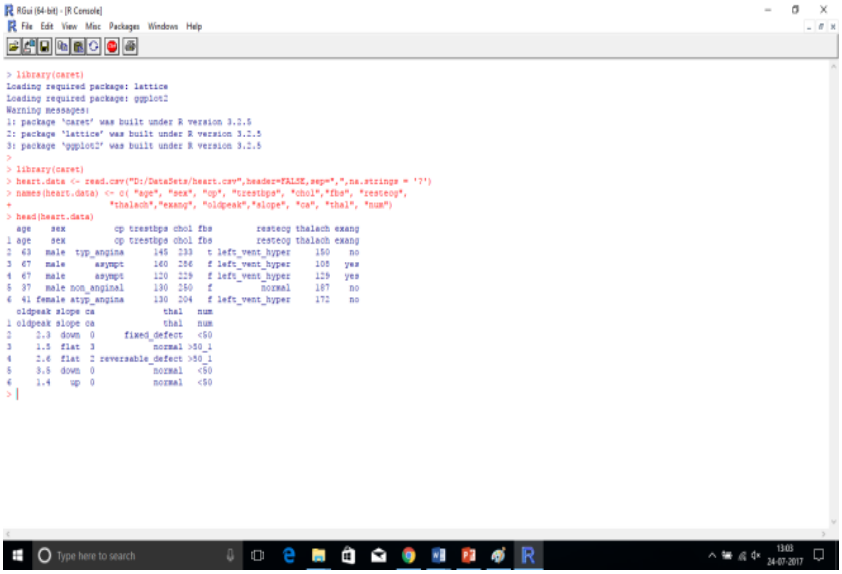

Figure 2: Read Dataset

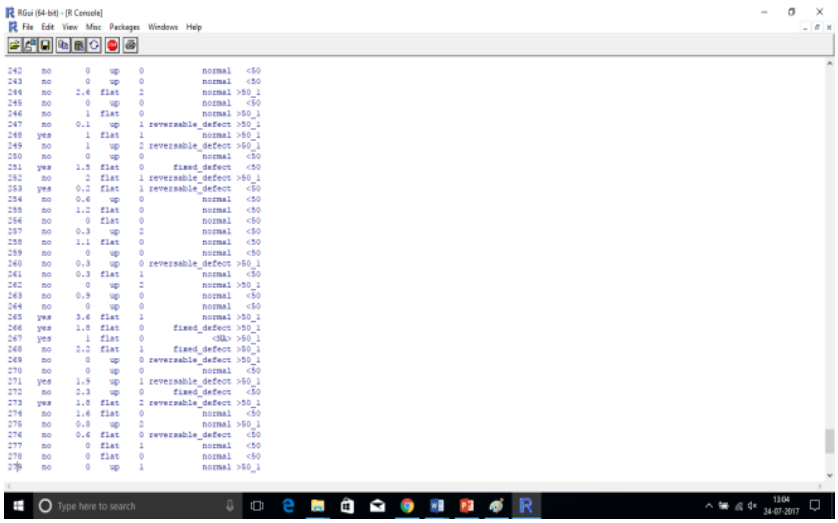

Figure 3: Actual Data

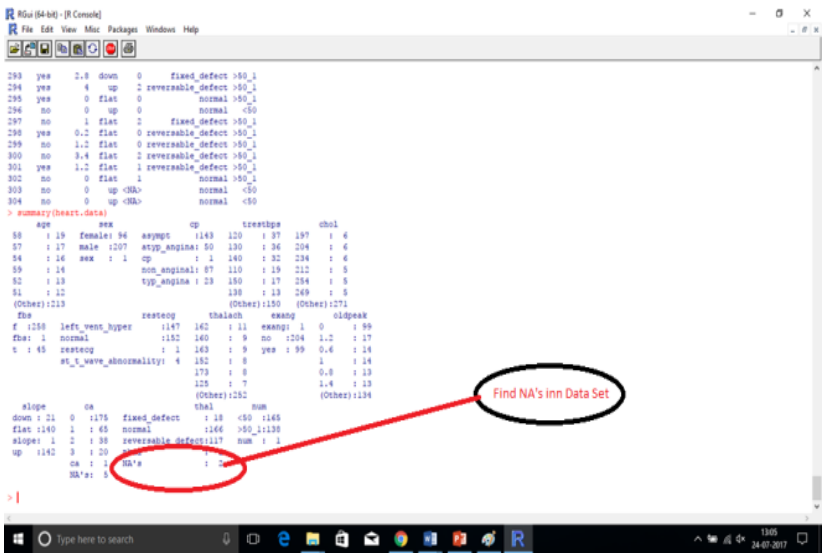

Figure 4: Summarized Data and finding the missing values

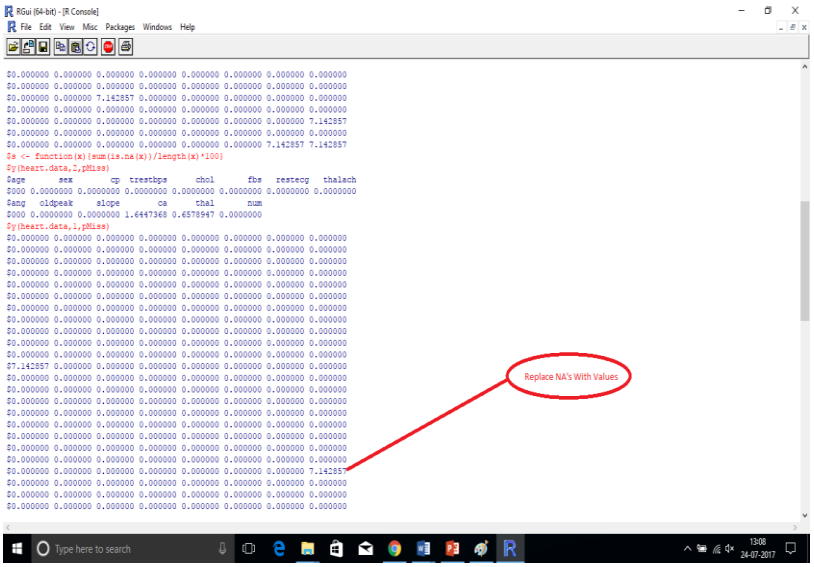

Figure 5: replaced by with their relevant values on the patient data by statistical approach.

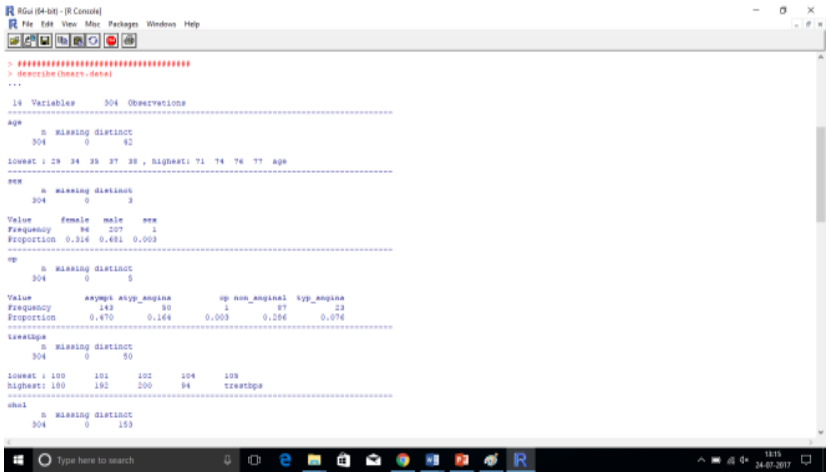

Figure 6:

```
describe <- function(...) {
      Hmisc::describe(...)
}
```

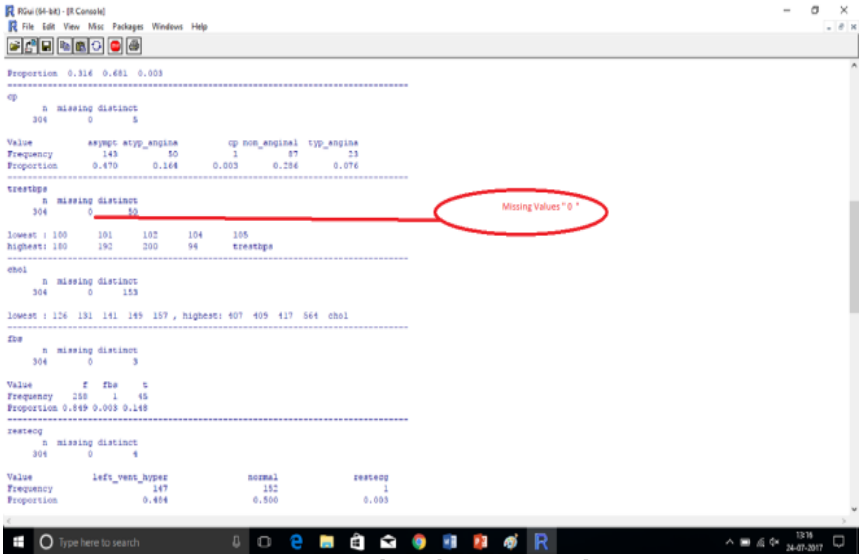

Figure 7 : replace the missing values.

Figure 8: Data Visualization

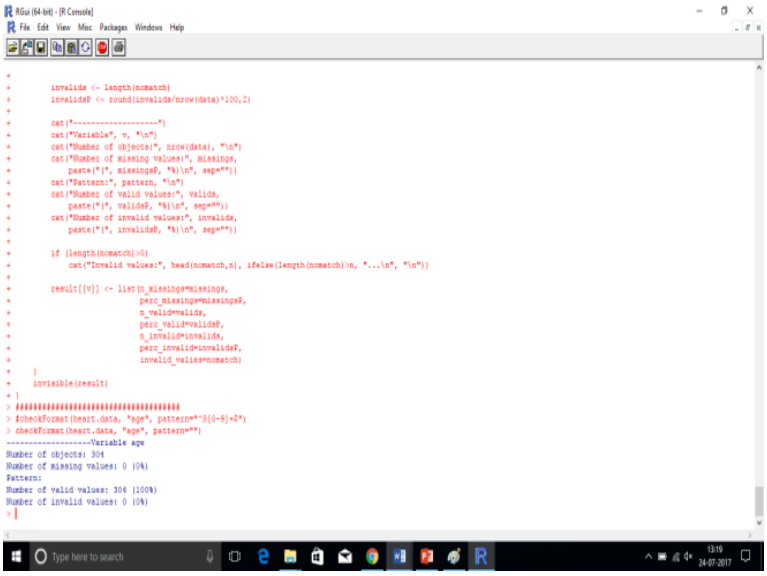

Figure 9: Accuracy (Authenticity)

Check Data Format

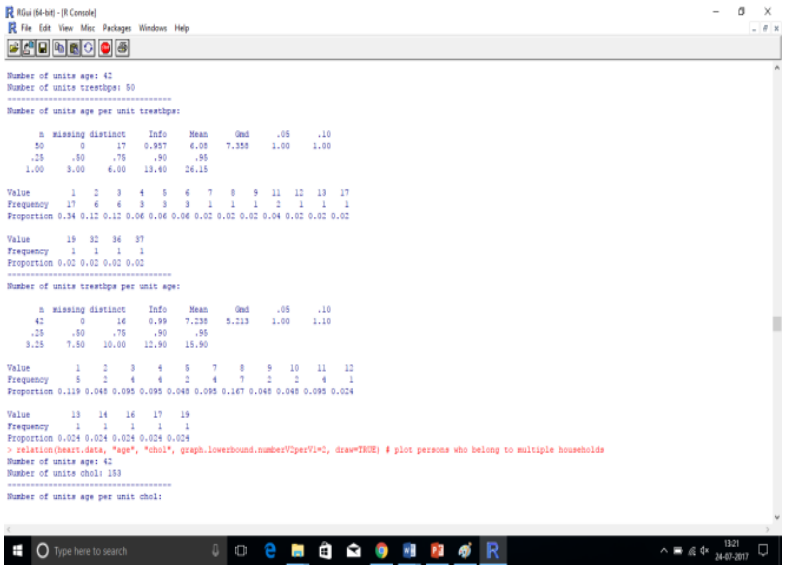

Figure 10: Inconsistent objects #overview of 1-to-n and m-1 relations between two variables.

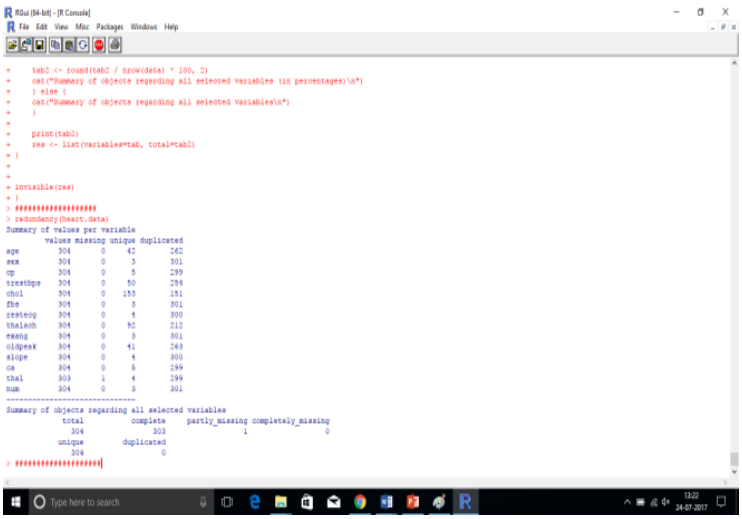

Figure 11: Result here by using my research we show that data was not qualified for the further analysis.

VI. CONCLUSION AND FURTHER WORK

This paper describes an opening research dedicated to analyze data quality issues in a critical domain as e-Health monitoring. We note that this is a first attempt to analyze data quality issues in this kind of applications, and naturally this aspect requires further investigation. For example, we observe that the quality criteria presented previously are the core of data quality approaches but they are not exhaustive. Since the most part of quality criteria depends on the specificities of the environment and the user requirements, we plan to include other perspectives of quality as Quality of Service (QoS) and context-awareness, especially linked to data collection level. We also consider necessary to model user and system quality requirements and associate them to optimal procedures, metrics and measures. One or several quality evaluation methods, algorithms and procedures must be correlated. For the definition and application of such procedures, we consider important to take into consideration the granularity of data to be evaluated, in order to control the volume of data to be processed, transmitted and communicated. Also, we are particularly concerned about: Which are the preferences of medical and industrial experts on quality criteria? At which level of granularity the experts consider interesting to ensure the quality of data; only considering one data or a dataset, or several datasets? At which frequency will this evaluation be performed, continuous or episodic? Etc. A survey with medical and industrial experts is being prepared.

FURTHER WORK: Enhance my research prediction of heart dieses and analysis on various methodologies and my proposed prediction model.

ACKNOWLEDGMENT

I would like to thank Dr.B.D.C.N. Prasad for his guidance and support for preparing this paper.

REFERENCES

[1]. Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for Data Quality Assessment and Improvement. ACM Computing Surveys, 41(3), Article 16.

[2]. Even, A., & Shankaranarayanan, G. (2009). Utility cost perspectives in data quality management. Journal of Computer Information Systems, 50(2), 127-135.

[3]. Haug, A., Pedersen, A., & Arlbjørn, J.S. (2009). A classification model of ERP system data quality. Industrial Management & Data Systems, 109(8), 1053-1068. doi:10.1108/02635570910991292

[4]. Jing-hua, X., Kang, X., & Xiao-wei, W.(2009). Factors influencing enterprise to improve data quality in information systems application —An empirical research on 185 enterprises through field studies. 16th International Conference on Management Science & Engineering, September 14-16, 2009, Moscow, Russia.

[5]. Piprani, B., & Ernst, D. (2008). A Model for Data Quality Assessment. Lecture Notes in Computer Science,5333, 750-759. doi:10.1007/978-3-540-88875-8_99

[6]. Watts, S. G., & Shankaranarayanan, A. E. (2009). Data quality assessment in context: A cognitive perspective. *DecisionSupportSystems*,48,202–211. doi:10.1016/j.dss.2009.07.012

[7]. D. Ratnam, P. HimaBindu, V. MallikSai, S. P. Rama Devi and P. RaghavendraRao, "Computer-Based Clinical Decision Support System for Prediction of Heart Diseases Using Naïve BayesAlgorithm",

International Journal of Computer Science and Information Technologies, vol. 5, no. 2 (2014) pp.2384-2388.
[8]. T. Santhanam and E. P. Ephzibah, "Heart Disease Prediction Using Hybrid Genetic Fuzzy Model", Indian Journal of Science and Technology, vol.8, no. 9, (2015)**,** pp.797–803.
[9]. G. Purusothaman and P. Krishnakumari, "A Survey of Data Mining Techniques on Risk Prediction: Heart Disease",Indian Journal of Science and Technology, vol. 8, no. 12, (2015).
[10]. S. J. Gnanasoundhari, G. Visalatchi and M. Balamurugan, "A Survey on Heart Disease PredictionSystem Using Data Mining Techniques", International Journal of Computer Science and Mobile Application, vol. 2, no. 2 (2014)**,** pp. 72-77
[11]. K. Lokanayaki and A. Malathi, "Exploring on Various Prediction Model in Data Mining Techniques for Disease Diagnosis", International Journal of Computer Applications, vol. 77, no. 5, (2013)**,** pp. 0975 – 8887.
[12]. A. Wilson, G. Wilson and J. Likhiya, "Heart Disease Prediction using the Data Mining Techniques", International Journal of Computer Science Trends and Technology (IJCST), vol. 2, no. 1, **(2014)**, pp.2347-8578.
[13]. Beant Kaur and Williamjeet Singh., "Review on Heart Disease Prediction System using Data Mining Techniques", IJRITCC ,October 2014.
[14]. Hlaudi Daniel Masethe, Mosima Anna Masetheprediction of Heart Disease using Classification Algorithms; Proceedings of the World Congress on Engineering and Computer Science 2014.
[15]. Vikas Chaurasia, et al, Early Prediction of Heart Diseases Using Data Mining Techniques; Caribbean Journal of Science and Technology ISSN 0799-3757, Vol.1,208-217, 2013.
[16]. Nilakshi P. Waghulde, Nilima P. Patil,"Genetic Neural Approach for Heart Disease Prediction", International Journal of Advanced Computer Research (ISSN (print): 2249-7277 ,Vol 4 Number-3 Issue Sept 2014.
[17]. Sudha A, Gayathiri P, Jaisankar N," Effective analysis and predictive model of stroke disease using classification methods", International Journal of Computer Applications. 2012; 43(14):26–31.
[18]. S.Suganya, P.Tamil Selvy, " A Proficient Heart Disease Prediction Method Using Fuzzy-Cart Algorithm, "IJSEAS, Volume2, Issue1, January2016.
[19]. B.V. Baiju and R.J.Remy Janet, "A survey on heart disease Diagnosis and Prediction using Naïve Bayes in Data Mining", IJCET, Vol 5, No.2, April 2015.
[20]. Rajwant Kaur, Sukhpreet Kaur, "Prediction of Heart disease Based on Risk Factors Using Genetic SVM Classifier", IJARCSSE Volume 5, Issue 12, December 2015.
[21]. Moloud Adbar, Sharareh R.Niakan Kalhori, tole Sutikno, Imam Much Ibnu Subroto, Goli Arji, "Comparing Performance of Data Mining algorithms in Prediction Heart Diseases", IJECE, Vol 5, No 6, December 2015.
[22]. S. Kiruthika Devi * , S. Krishnapriya and Dristipona Kalita Prediction of Heart Disease using Data Mining TechniquesIndian Journal of Science and Technology, Vol 9(39), DOI: 0.17485/ijst/2016/v9i39/102078, October 2016.
[23]. V. Krishnaiah Heart Disease Prediction System using Data Mining Techniques and Intelligent Fuzzy Approach: A Review International Journal of Computer Applications (0975 – 8887) Volume 136 –No.2, February 2016.
[24]. Syed Immamul Ansarullah Heart Disease Prediction System using Data Mining Techniques: A study international Research Journal of Engineering and Technology (IRJET) Volume: 03 Issue: 08 | Aug-2016.
[25]. Sonam Nikhar "PREDICTION OF HEART DISEASE USING DATA MINING TECHNIQUES" - A Review International Research Journal of Engineering and Technology (IRJET) Volume: 03 Issue: 08 | Aug-2016.

*K.Parish Venkata Kumar . "Effective Data Quality Management In Health Sector." International Journal of Engineering Research and Development, vol. 13, no. 09, 2017, pp. 46–54.