International Journal of Database Theory and Application
Vol.8, No.3 (2015), pp.333-342
http://dx.doi.org/10.14257/ijdta.2015.8.3.29

Enhanced Extraction Clinical Data Technique to Improve Data Quality in Clinical Data Warehouse

AbubakerElrazi O. Mohammed[1] and Samani A. Talab[2]

[1]*Department of Computer Science, Shendi University, Sudan*
[2]*Department of Computer Science, Elnileen University, Sudan*
razi190@gmail.com,profsamani@gmail.com

Abstract

ETL process represents a major part in the process of clinical data warehouse development, where the efficiency of DWH is mainly depending on ETL component and its architecture. In medical field there are a huge clinical data stored in several medical operational systems during receiving medical services. However, extracting of these data are complex, time consuming, and labor intensive task to ensure high data quality before all kinds of data analyses. Moreover, integration of clinical data from various sources is challenges; where these data have been integrate from heterogeneous data sources from multiple health institutions with incompatible structures. Furthermore, heterogeneous clinical data are stored dispersed and isolated from one another. Thus, these clinical data need to be extracted and integrated into the clinical data warehouse through a robust extraction technique. This paper introduces an enhanced ETL technique, which integrate clinical data form heterogeneous data source into staging area.

Keywords*: ETL, Data Warehouse, Clinical Data Warehouse, Data Integration*

1. Introduction

The clinical data are stored in various information systems, such as hospital information system, radiological information system, laboratory information system, and picture archiving and communication System [1]. However, the clinical practices and their routines in different hospitals are different significantly, as reported in [2]. Moreover, medical data is dispersed throughout the medical systems, being stored in proprietary systems with incompatible architectures. Furthermore, the process of extracting the medical data from these various information systems is time-consuming and labor intensive [3]. Therefore, the integration process of medical data from various systems into CDWH is sensitive process in order to deliver quality patient care. Additionally, the confidentiality and privacy of the medical data must be observed.

The Clinical Data Warehousing (CDWH) is integrating data from various operational medical and administrative systems into one common data source, in an efficient and secured way, which is optimized for intelligent data analysis purposes [4]. Therefore, the CDWH provide information to users in areas ranging from research to management [5, 2]. The CDWH facilitate efficient storage, enhances timely analysis and improves the quality of real time decision making processes and timely process [6,7].

Consequently, extraction technique is developing to integrate clinical data and improve the quality of data in CDWH. Where, the extraction technique is responsible for the extraction and integration of data from several sources. The efficiency of ETL processes is mainly depending on extraction process. Figure 1 shows the general framework for ETL processes.

ISSN: 2005-4270 IJDTA

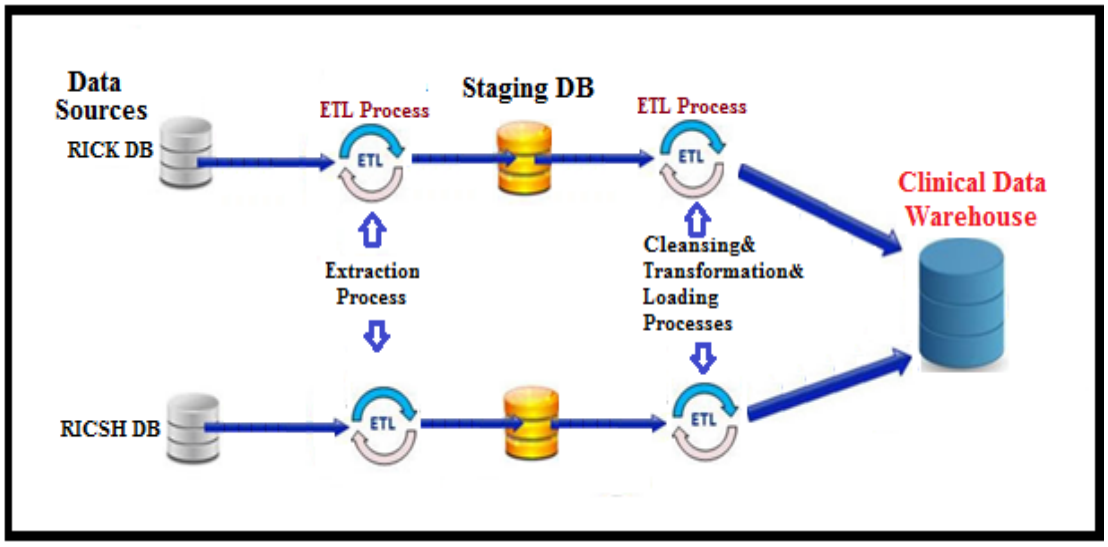


Figure 1.Processes of the ETL

The presentation of the study is organized as follows. Section 2 review related work in ETL. Then, Section 3 is discussing the extraction process issues. Section 4 is discussing the implementation of the extraction process. Section 5 the discussion. Finally Section 6 concludes the study.

2. Related Work

This section provides brief overview of some previous work efforts related to the work in this study. Michael Blechner *et al*. [8], proposed extraction algorithms, and establishing a working CRDW base on a star schema. The proposed star schema preserves the contextual and semantic information from the CDA R2 documents by designing a CDA Statement fact table and supporting it with new/reused dimension tables to form a star. The CDAW Statement Fact Table is structured to capture the details of each clinical statement encoded using CDAW clinical statements. Xishui Pan, *et al*. [9], have introduced a technique to integrate the heterogeneous clinical data sources and support the direct copy from these data sources to ODS database by ETL process. This study aims to improve ETL process performance. The proposed ODS data model consisted of two components include ODS and two-step ETL subcomponents to perform the ETL tasks. Moreover, ETL task has been separated into two core steps in enhanced ETL component: (1) dynamic filter and copy of the original operational data sources to ODS; (2) specialized transforming the ODS data to detail clinical data warehouse. Zehai Li, *et al*. [10], have proposed a novel conceptual model based on CommonCubes, for the modeling of ETL processes by providing formal definitions and descriptions of essential ETL entities. ETL entities include data source, data target, ETL function, and ETL mapping from source attributes to target attributes to support the design of ETL processes. CommonCube described the schemas of data cubes, which is kept compatible with various data warehouse models, and it can release the design of ETL processes from overdependence on the target data warehouse. The proposed ETL process divided into two phases include: (1) Design phase; in this phase, the designer defines the schema mappings from the data sources to the target data warehouse, and stores them into files (or DBMS) as ETL tasks. (2) Executing phase; in this phase, ETL tasks defined in design phase are executed according to the schedules made up by the designer to perform extracting, cleansing and loading data from the data sources to the target data warehouse.

These proposed ETL techniques and approaches aim to develop ETL system to support the development of CDWH. But they do not describe how to efficiently dealing

with data integration issues (such data extraction issues). Furthermore, these proposed structures do not discuss how to efficiently dealing with data quality issues.

3. The Extraction Process Issues

The extraction process is responsible for extracting relevant data from heterogeneous data sources. Where, the ETL process requires connection to the source systems, and selecting the relevant data needed for analytical processing within the CDWH. However, the data extract from numerous disparate source systems and each of these data sources has its distinct set of characteristics that need to be managed in order to effectively extract data for the ETL process. Furthermore, the complexity of the extraction process depends on the data characteristics and attributes, amount of source data and processing time. Therefore, the ETL process needs to effectively integrate technology to extract these data. Handling extraction process issues and challenges need to provide the following requirements to ensure subject-oriented of the CDWH as reported in [11]:

1. Analyzing data sources in order to comprehend their structure and contents to understand the datathat exist in the databases to identify the relevant data at these sources that depending on the purpose of CDWH, the selection of these data requires:
    a. Identifying source systems that contain the required data and identifying the quality and scope of each data source.
    b. Understanding the format of data stored by each source to determine whether all the data available to fulfill the requirements or not, and the required data fields populate properly and consistently.
    c. Identifying the attributescontain in each data source.
2. Determining the options of extracting the data from the source systems which include updated notification, incremental extracts, and full extracts to capture only changes in source files.
3. Determining the protocols for data transferring.
4. Determining encryption standards need to be set with each of the source systems.
5. Monitoring data transfer failures and errors and making notifications through different methods such as control files, metadata files, email notifications, system log writing and file system log writing.

The proposed Extraction technique consist three sequential stages. These stages include; medical analysis, the physical design of process, and ETL extraction process evaluation. Furthermore, the medical analysis process is important issue in order to study and analyze the existing process from medical perspective as well as to determine requirements, and the requirements are further analysis and investigate to determine the data integration and quality data problems. These requirements include: clinical data requirements and clinical data integration requirements, and clinical data quality requirements. The extraction technique designing base on these requirements, in order to maintains data integration and improve data quality. The clinical data integration and quality problems will be handled within extraction technique. These clinical data integration and quality problems classification are summarized in Table 1.

Table 1. Clinical Data Quality and Clinical Data Integration Issues Classification

Data Quality Problem	Problem Description	Most the possible causes of data quality Problems
Medical Purpose Problem	The medical purpose and requirement is not determined in proper ways.	- Incomplete or wrong requirement analysis of the project leads to poor schema design. - Lack of currency in medical rules cause poor requirement analysis which leads to poor schema design.
Identify Problem	Inadequate selection of data sources required to achieve medical purpose.	- Sources which do not comply with medical rules. - Different medical rules of various data sources. - Usage of decontrolled applications and databases as data sources for CDWH in the organizations.
Relevancyproblem	The data collected is not relevant to the medical purpose.	- Inadequate selection of relevant data from selected data sources.
Loss of data Problem	The data is loss during the process of transferring of data form source to staging area	- Loss of data during the extraction process (rejected records).
ScalabilityProblem	An ETL process is not able to handle higher volumes of data.	- Multiple data sources generate semantic heterogeneity. - As time and proximity from the source increase, the chances for getting correct data decrease.
AvailabilityProblem	The ETL process is not operational during a specific time period, in other word the data are available when required	- The resources of the system that needed are not available when needed. - The required data in the data warehouse is not available within specified time and accuracy constraints.
Completenessproblem	All the requisite information is not recorded /available, or in an unusable state (the data isn't thorough in the attributes that require them).	- Fields with null values. - Some fields with false or incomplete values.
Consistencyproblem	The data is not satisfies a set of constraints, and not maintained in a consistent fashion. Data values doesn't consistent across data sets.	- Inconsistent use of special characters.
DWH required format Problem	Data in inappropriate forms for mining.	- Different data types for similar columns (A patient ID is stored as a number in one table and a string in another). - Fields with Inconsistent/Incorrect data formatting (E.g. specific attribute is stored in one table in the specific format and in another table in different format).

The Extraction technique will design and develop to address the following questions about the problems that affect the data integration and the data quality during extraction process.

1. Is the medical rules determine in proper ways?
2. What are the data extraction problems that affect the quality and integration of data?
3. Are you using sources in complying with the medical rules?
4. Do you have multiple formats to be accessed- relational DBs, flat files, *etc*.?
5. How data extraction process techniques search the relevant clinical data?
6. What are the transferred options of the extracting data from the systems into staging area (update notification, incremental extracts, or full extracts)?
7. What are the required/available frequencies of the extracts?

4. Implementationof Extraction Process

Analysis of extraction process requirements aim to determining and defining data extraction problems that may affect the quality of data and integration process, examining how to select the relevant data, and how to transfer these data to staging area. The data problems include clear understanding of medical purposes, identifying the required data, scalability, consistency, integrity, completeness and format problem. However, the required data stored in several data sources which are different structure and technologies. This required clear understanding of these sources structure and technologies to understand the format and attribute of data stored by each source. Where, identifying of the data sources depend on the purpose of the developing of CDWH.

The required data categorize in three types of medical data related to the patients. These categories involve cancer disease information, demographic information, and patient clinical records information. The first category, cancer disease information refers to general cancer's information, such as cancer types, cancer stages, diagnosis types, symptoms, treatment types, risk factor types, physicians and *etc*. The second category, demographic data refer to personal patient's information such as location, occupation, sex, education, parent's relatives, tribe and *etc*. The third category, patient clinical information refers to transaction data that collected about each patient during having treatment. These data involved patient diagnosis, patient treatment procedure, treatment Result, treatment side effects, laboratory test and *etc*.

The extraction process consists of two phases, initial extraction, and incremental extraction. In the initial extraction, the relevant clinical data extract from data sources for the first time. This process is done only one time after developing the CDWH. On the other hand, the incremental extraction refreshes the CDWH with the modified and added data in the data sources since the last extraction process. This process is done periodically according to the medical needs. Once data extracted from source systems according specific rules and conditions, the data are transferred to staging area. Furthermore, the transfer process is monitoring and making notifications when failure and error occur. In CDWH update process ETL technique captures only change in data sources since the last extraction process.

Algorithm: Data Extraction Process

i. **Identifying and analyzing the data sources and creating the data sources list for each data source:**
   1. Identifying the list of data sources that contain the required data according to medical needs to observe medical goals,
   2. Indentifying the type of the databases (format of data stored by each source),
   4. Analyzing each of data source to identifying the structure of data sources (format of data stored by each source).
   5- Identifying the attributescontain in each data source,
   6. Checking if it is a new source add to the data source list or checking if there any object added to data source.

ii. **Establishing connection and extracting data:**
   7. Determining the type of the data source,
   8. Using appropriate drivers to establish the connection,
   9. Identifying the data source object that contains the required data.
   10. Select the relevant data according medical objectives;
   11. Checking to select the required extraction option (initial/ or incremental) to perform the required processes.
   12. Mapping the data source and data staging area schemas,
   13. Identifying the data problems that affect the data quality at extraction process and handle these problems with appropriate method.
   14. Indentifying the options of extracting data from the source systems into staging area,
   15. Identifying the appropriate extraction strategy.

iii. Loading of extracted data into data staging are:

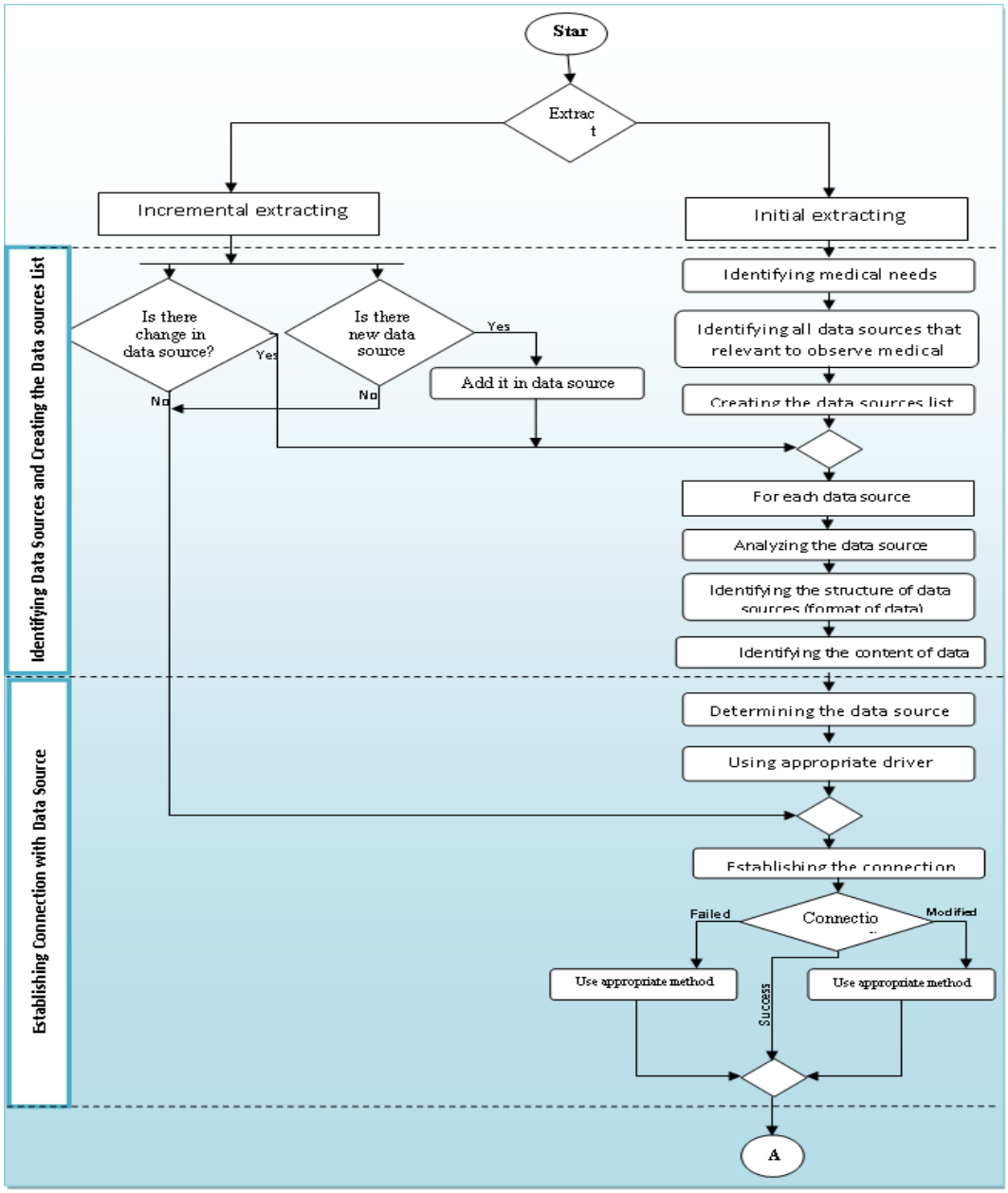
16. Establishing connection with data staging area using appropriate data connection for transferring data,
17. Transferring the data into data staging area.

iv. Modification /updating of data Staging Area:

18. Identifying the changes in the data sources,
19. Update DSA.

v. Monitoring data transferring:

20. Monitoring data transfer failures and errors,
21. Making notifications.



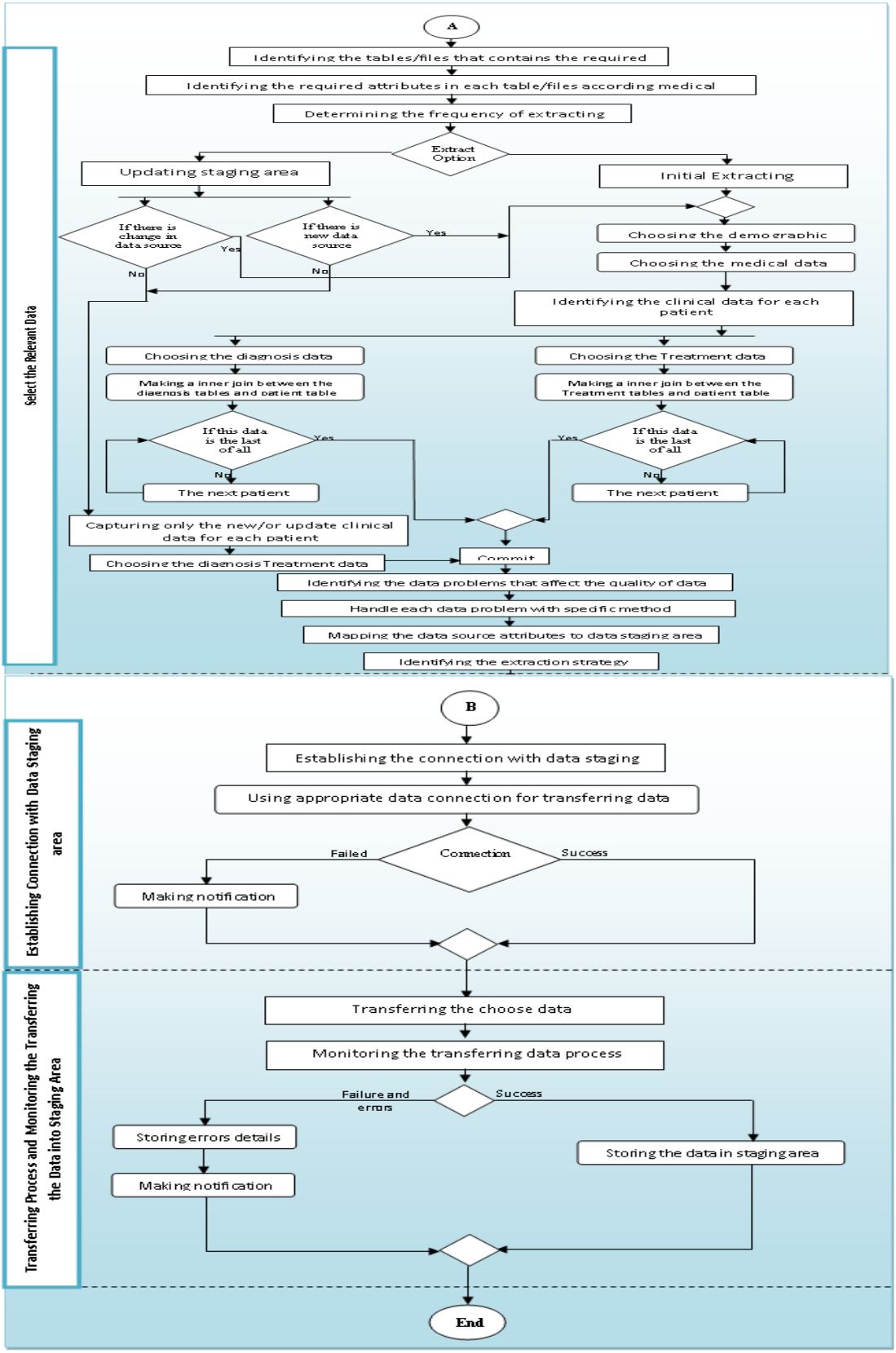


Figure 2. Flowchart of Data Extraction Technique

5. Discussion

Medical field is more complicated than the business area and produces a set of integration issues and challenges. These issues involve clear identification of extracting and integrating requirements as well as developing and evaluating the extraction technique. These requirements include the clinical data requirements, clinical data integration requirements, and clinical data integration requirements. However, the clinical data is different from business data where the clinical data produce new requirements, if not considered the quality of data will be affected. The complexity of the hospital environment evolves the diagnosis and treatment procedures and their relation with other information produced distinct set of data characteristics. Moreover, the data collected from different hospitals which used and diverse data format and DBMS. These complexity of clinical data rise several issues such as; poorly characterized mathematically, difficult data type for mining, and difficult to determine hidden relationships.

The functions of extraction process include: the identification of relevant data at the various data sources, the extraction of these data, and transportation of these data to the stage area. Whereas, clear understanding of the medical purpose represents an important issue in the process of developing extraction technique.

Consequently, the study discussed how the clinical data extraction and integration data are performed. The medical data requirements are collected to understand the problems domain in addition to determine the suitable solution to handle these problems. Furthermore, extraction technique aims to integrate large volumes of data collected from several clinical information systems. Therefore, development of effectively integrated systems that have different platforms is required. The extraction technique is proposed and designed through four sequential phases which include: medical analysis, physical development, and evaluation. Finally, the produced data will be evaluated against acceptance criteria to ensure that the medical objectives are achieved. The data quality assessment process includes:

- Evaluation of extracted data pattern is performed to identify the truly interesting patterns representing knowledge.
- Establishing metrics to assess the validity of the data extracted from the various systems.
- Systematically reviews all the data elements, considering factors such as ranging of values, number of null records, duplicate records, compliance with medical rules, and inaccurately recorded information.
- Providing a summary of data problems and a strategy to handle these problems.

6. Conclusion

This paper proposed the algorithm of develops successful clinical data extraction. The usage of DWH technologies in medical field produced new issues to DWH technologies. Handling these issues requires determining the requirements of extraction system that include: clinical data requirement, data integration requirements, and data quality requirements. Furthermore, extraction process required a clear definition of clinical purpose to determine these requirements, and build a robust technique to integrate data and handle all data quality problems at extraction process. This paper presented designing of extraction algorithm in order to integrate clinical data from disciplined medical sources. All required medical data are gathered from a variety of clinical and administration sources and merged into the staging area.

References

[1] S. T. Wong, K. S. Hoo, R. C. Knowlton, K. D. Laxer, X. Cao, R. A. Hawkins, "Design and applications of a multimodality image data warehouse framework," Journal of the American Medical Informatics Association, vol. 9, **(2002)**, pp. 239-254.
[2] T. R. Sahama and P. R. Croll, "A data warehouse architecture for clinical data warehousing," in Proceedings of the fifth Australasian symposium on ACSW frontiers-Vol. 68, **(2007)**, pp. 68.
[3] W. Reed, S. Jor, and R. Bjugn, "How can clinical biobanks and patient information be adapted for research–Establishing a hospital based data warehouse solution," Norsk epidemiologi, vol. 21,**(2012)**.
[4] J. Widom, "Research problems in data warehousing," Proceedings of the fourth international conference on Information and knowledge management, **(1995)**, 25-30.
[5] A. Sen and V. S. Jacob, "STRENGTH," Communications of the ACM, vol. 41, **(1998)**, pp. 29.
[6] G. Purusothaman and P. Krishnakumari, "Hybrid Model for Clinical Diagnosis and Treatment Using Data Mining Techniques," The International Journal Of Engineering and Science (IJES), vol. 3, **(2014)**, pp. 39-42.
[7] N. Esfandiary, M. R. Babavalian, A.-M. E. Moghadam, and V. K. Tabar, "Knowledge Discovery in Medicine: Current Issue and Future Trend," Expert Systems with Applications, vol. 41, **(2014)**, pp. 4434-4463.
[8] M. Blechner, R. K. Saripalle, and S. A. Demurjian, "A proposed star schema and extraction process to enhance the collection of contextual & semantic information for clinical research data warehouses," Proceedings of the IEEE International Conference, Bioinformatics and Biomedicine Workshops (BIBMW), , **(2012)**, 798-805.
[9] X. Pan, X. Zhou, H. Song, R. Zhang, and T. Zhang, "Enhanced data extraction, transforming and loading processing for traditional Chinese medicine clinical data warehouse," Proceedings of the 14th IEEE International Conference, e-Health Networking, Applications and Services (Healthcom), **(2012)**, 57-61.
[10] Z. Li, J. Sun, H. Yu, and J. Zhang, "Commoncube-based conceptual modeling of ETL processes," Proceedings of the Fifth ICCA International Conference in Control and Automation, **(2005)**, 131-136.
[11] R. O. Mohammed and S. A. Talab, "Clinical Data Warehouse Issues and Challenges," International Journal of u-and e-Service, Science and Technology, vol. 7,**(2014)**, pp. 251-262.

Authors

Abubaker Elrazi Osman Mohammed,he obtained his BSc degree in computer & Statistical Science from Gazira University-sudan in 2000. He received his MSc in computer science from Gazira University- sudan in 2006. He received PhD degree in Shendi University-Sudan in 2015. His research interest is data warehouse and data mining. He is working as assistance professor atShendi University-Sudan.

Elsamani Abd Eltalab,he received BSc, MSc and PhD degree in computer science from department of computer science, University of Khartoum, Sudan in 1989, 1995, and 2001 respectively. Currently, he is working as professor at Faculty of Computer Science and Information Technology, AL_Neelain University, Khartoum, Sudan. His fields of interest are in data structures, algorithms, teaching and learning, compiler design and numerical computation.