DATA WAREHOUSE FOR MINING AND SIMULATING MEDICAL RECORDS

[1]*FRED FUDAH MOVEH, [2]MUHAMMAD ABBA JALO, [3]ABDULAZIZ SULEIMAN YAHYA and [4] IBRAHIM SHOK

[1, 2&3]Department of information Technology, Modibbo Adama University, Yola

[4] Transmission Company of Nigeria (TCN)

Corresponding Author: fredmoveh@mau.edu.ng

ABSTRACT

Nowadays, patient's data required to make informed medical decisions are trapped within fragmented and disparate clinical and administrative systems that are not properly integrated or fully utilized. Therefore, there is a growing need in the healthcare sector to store and organize sizeable record of patients to assist the healthcare sector in making informed decision. This study has been able to design and simulate the framework by developing a web application that integrates the front-end, middle-end and the back-end together as a single system using patient records from three different tropical diseases; Malaria, Measles and Typhoid Fever as the operational data source for populating the Data Warehouse. The application can be implemented in any hospital in other to aid quick decision making, most especially in certain cases for community diagnosing. Previous studies only focused on specific diseases and also visualization of clinical work flow. However, this research provides an architecture for designing a clinical data warehouse that is not limited to a single disease and will operate as a distributed system. Periodic update on a daily basis is possible because contemporary technologies have narrowed the gap between updates which enable organizations to have a "real time" data warehouse which can be analysed using an OLAP Server. In conclusion, the study demonstrates how data can be incorporated from diverse desperate heterogeneous clinical data stores into a single data warehouse for mining and analysis purpose to aid medical practitioners and decision makers in decision-making. The study has also been able to develop data mining, reporting and analysis tool (GUI) where users can interact with the system to get a speedy and timely information needed for the clinical decision making and community diagnosing.

Keywords: Data warehouse, Data mining, Hospital records

INTRODUCTION

Analysis of medical record is globally an indispensable tool in health information. Health data is more than just statistics. It can be controlled, used and shared in so many different ways. However, ignoring certain medical data has the potential to change the way a patient is treated, how care is provided and what happens to the patient. As the adoption of medical record as a data source increases, novel methods in biostatistics for analysing health record are needed to drive utility in the form of clinical findings (Swinton et al., 2018).

Jothi *et al*., (2015)defined medical record as a confidential record that is kept for each patient by healthcare professional or organization. It contains patient's personal details (such as name, address, date of birth), a summary of patient's medical history and documentation of each event, including symptoms, diagnosis, treatment and outcome. These information's obtained from analysis of medical records provide the essential data for monitoring patients care,

clinical audits and assessing patterns of care and service delivery (Faramarz, Hossein, Alireza &Johan, 2018). It is in the consideration of these that this study focuses on data warehousing in the management of patient records to improve decision making. The Specific objectives of this study include:

i.   Design a database for the Patients Clinical Data
ii.   Simulate the data from the database in order to generate reports from the clinical record to aid decision making using Microsoft SQL Server Queries and SQL Server Integration Services (SSIS).
iii.   Develop a web platform that will integrate the interface with the back-end using an object oriented programming concept.

RELATED STUDY

Abubakar et. al., 2014, conducted a study which focused on the use of OLAP and ETL technology in designing Diabetes data warehouse. However, the implementation tools were made on Microsoft SQL Server 2008, SQL server integration service 2008 (SSIS), SQL server analysis service 2008 (SSAS), SQL server Reporting service 2008 (SSRS) and C # language.

The diabetes data warehouse allows the user to analyse the diabetes, estimate the cost of treatment, measure the impact of a particular drug on the disease and identify death rate with respect to the type of diabetes. It also enables healthcare providers to detect and subsequently diagnose problems earlier and provide support for informed healthcare decisions. It assists the healthcare providers in improving care for patients by identifying blood sugar levels and blood pressure to manage treatments. It provides patients with ways of caring for themselves by recommending and monitoring food consumption, physical activity and insulin dosage. The study concludes that the data warehouse will assist executive managers and doctors in providing accurate and timely information for Healthcare decision making. Another study conducted by Vankipuram et al. (2018) used Radio Frequency Identification (RFID) data collected to develop pertaining to clinical workflow. The study presents a set of analytics derived from sensor-based automated location tracking (capturing individual movement patterns as well as interaction patterns), that can be used in an exploratory capacity by researchers and clinicians as well as to complement existing qualitative techniques. . The aim was to develop analytic techniques with interactive visualizations which can be used in safety and efficiency of care offered in an exploratory capacity to identify areas of concern and to potentially supplement qualitative methods in the future.

Another study was carried out by Huang et. al. (2015), the objective of their study was to create a visually interactive exploratory data analysis tool that can be used to graphically show disease associations over time. That is, the tool presents how a group of patients with one chronic disease may go on to develop other diseases over time. The researchers developed a standardized data analysis process to support cohort study with a focus on a particular disease.
Bum et. al. (2019) tried to tackle the problem of interpretability and interactivity by designing a visual analytics solution with Recurrent Neural Networks (RNN) based model for predictive analysis tasks on EMR data. The researchers' task was to predict the risk of a patient's future diagnosis in heart failure and cataract, based on information from previous medical visits in the EMR dataset. Their study involved iterative design, assessment and discussion activities

between medical experts, artificial intelligence scientists and visual analytics researchers.

Swinton et. al. (2018) studied a statistical framework for interpreting individual response to intervention: paving way for personalized nutrition and exercise prescription. The researchers described procedures required to interpret data collected from individuals both pre- and post-intervention.

MATERIALD AND METHODOLOGY

To develop the data warehouse for simulating medical record, a system architecture for Data Warehouse and Mining Clinical Records with the ability to Integrate the independent and dependent Data Marts within the architecture, has a multi-tier Extract, Transform and Load (ETL) which are simpler and in different stages, has a standard access points for all medical and non-medical experts (Using a three tier server), and the ability to Mine and analyse records using Data Visualization was develop.

The architecture has a back and frond end system in which so many activities are carried out. The back end systems comprise of the operational data source system, data staging area and the data presentation area. Data are first extracted from different operational data source systems using ETL and then stored at the data staging area where it is being processed as soon as it is captured. The activities of the ETL at the data staging area include data cleansing and validation, data integration, data fixing and data entry errors removal, transforming and refreshing data into a new normalized standard. As soon as data is cleaned, the transformed data are loaded and indexed into the data presentation area where the Data Warehouse (DW) is located. Figure 1.1 illustrates the general system architecture for

this study. It shows how the Source System database, Servers and the User/administrator system are connected for possible data visualization and decision making.
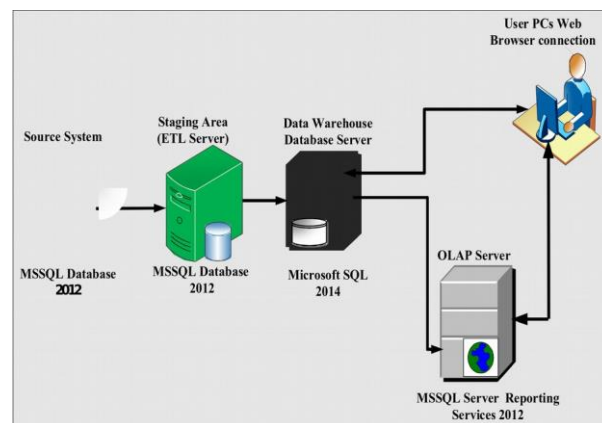

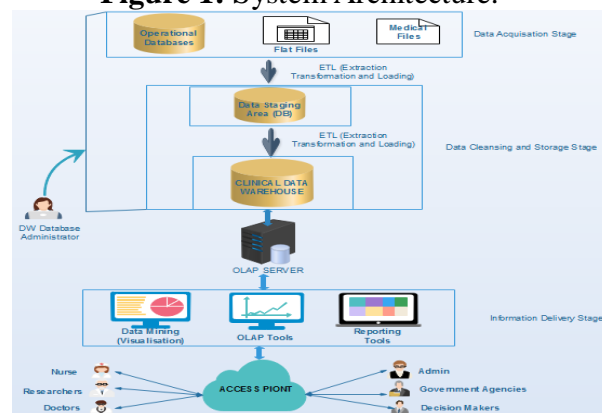**Figure 1:** System Architecture.

Figure 2: Data Warehouse and Mining Architecture.

Microsoft Visio was used to design the logical Data Mart while the actual physical Data Mart Models of the required Data Marts on an SQL Server Database Management System 2019 were designed. The physical ETL design for the data load from the source system to the staging database was designed using SSIS and the data warehouse database was created on an MSSQL Server 2019 database, with the data loading also done through the SSIS package.

RESULTS

A system output is the most important component of a working system because the interactivity of the system depends on its output. This is the main reason why the output of a Data Mining and Analysis system determines the effectiveness and efficiency of the system. The systems' output presents information from records of three different tropical diseases; Malaria, Measles and Typhoid Fever as the operational data source for populating the Data Warehouse.

Analysis based on disease, month and year of diagnosis (for Malaria)

Figure 3: Malaria cases based on Month and Year

Analysis based on patients' marital status and year of diagnosis (for Malaria)

Figure 4: Malaria cases based on marital status and year of diagnosis

Analysis based on patients age group and year of diagnosis (for Malaria)

DIAGNOSED ILLMENT NAME	AGE	YEAR OF DIAGNOSIS	NUMBER OF PATIENTS
MALARIA	25	2020	6
MALARIA	26	2020	5
MALARIA	27	2020	4
MALARIA	28	2020	5
MALARIA	29	2020	5
MALARIA	25	2021	1
MALARIA	27	2021	2
MALARIA	29	2021	3
MALARIA	30	2021	1
MALARIA	28	2022	1
MALARIA	29	2022	3
MALARIA	30	2022	2

© 2023 Jalo's

Figure 5: Malaria cases based on patient age group and year of diagnosis

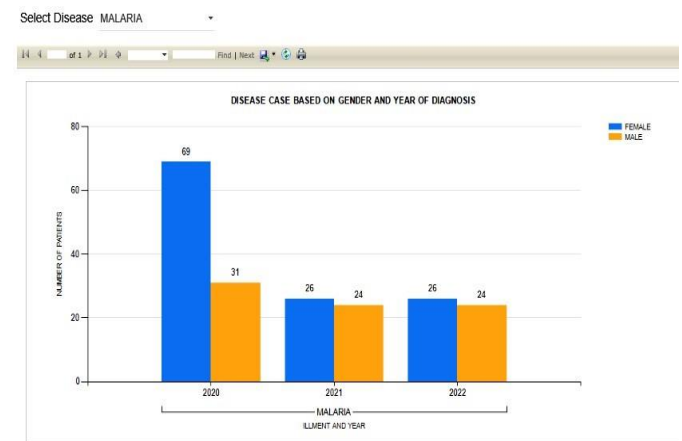## Data Visualization based on disease, gender and year (for Malaria)

© 2022 Jalo's

Figure 6: Malaria report based on patient gender and year of diagnosis

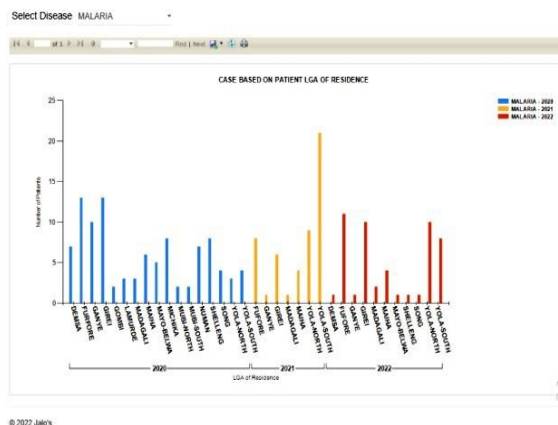**Data visualization based on LGA of residence (for Malaria)**

Figure 7: Malaria report based on LGA of residence and year of diagnosis

Data visualization of death rate based on marital status and year of diagnosis (for Malaria)
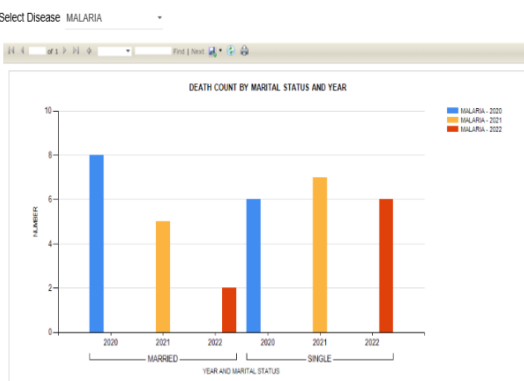

Figure 8: Malaria Death cases based on marital status and year of diagnosis

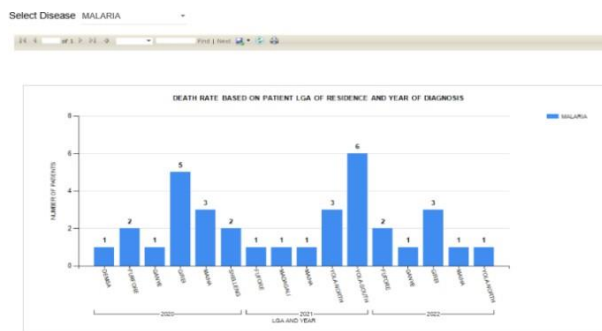**Data visualization of death rate based on patients' LGA of residence (for Malaria)**

Figure 9: Malaria death report based on LGA of residence and year of diagnosis

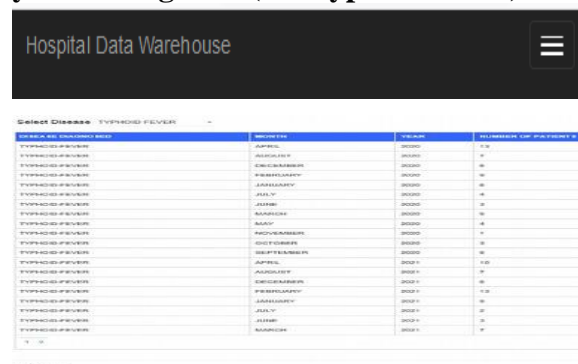**Analysis based on disease, month and year of diagnosis (for Typhoid Fever)**

Figure 10: Typhoid Fever cases based on Month and Year

Analysis based on patients' marital status and year of diagnosis (for Typhoid Fever)

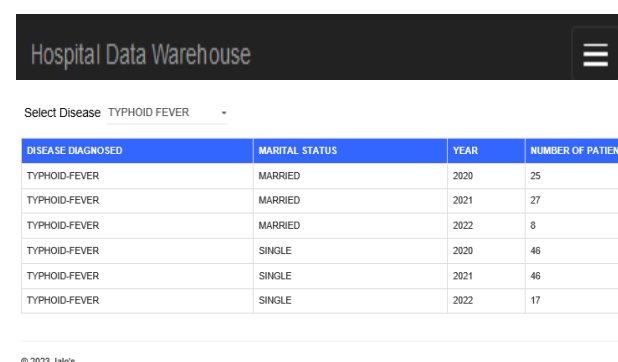| DISEASE DIAGNOSED	MARITAL STATUS	YEAR	NUMBER OF PATIENT
TYPHOID-FEVER	MARRIED	2020	25
TYPHOID-FEVER	MARRIED	2021	27
TYPHOID-FEVER	MARRIED	2022	8
TYPHOID-FEVER	SINGLE	2020	46
TYPHOID-FEVER	SINGLE	2021	46
TYPHOID-FEVER	SINGLE	2022	17

© 2023 Jalo's

Figure 11: Typhoid Fever cases based on marital status and year of diagnosis

Analysis based on patients age group and year of diagnosis(for Typhoid Fever)

Figure 12: Typhoid Fever cases based on patient age group and year of diagnosis

Data Visualization based on disease, gender and year (for Typhoid Fever)
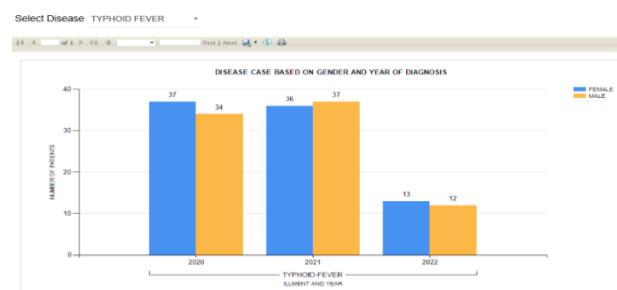
Figure 13: Typhoid Fever report based on patient gender and year of diagnosis

Data visualization based on LGA of residence (for Typhoid Fever)

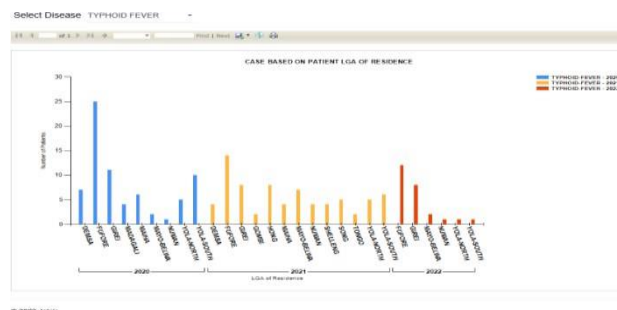**Figure 14:** Typhoid Fever report based on LGA of residence and year of diagnosis

Data visualization of death rate based on marital status and year of diagnosis (for Typhoid Fever)
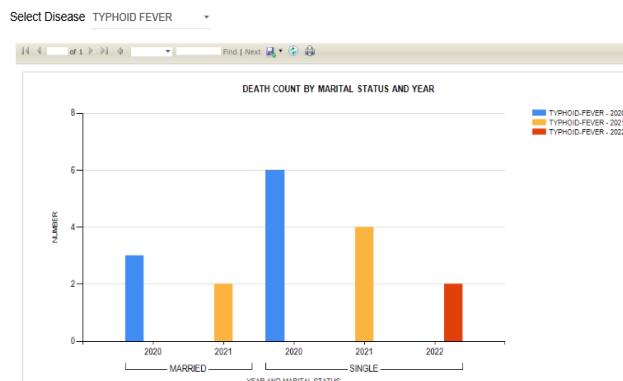
Figure 15: Typhoid Fever Death cases based on marital status and year of diagnosis

Data visualization of death rate based on patients' LGA of residence (for Typhoid)
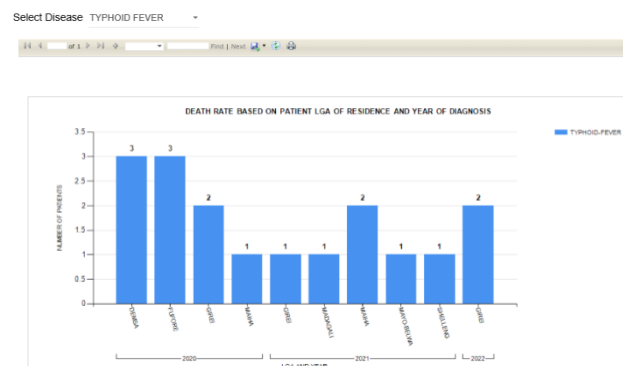
Figure 16: Typhoid Fever death report based on LGA of residence and year of diagnosis

DISCUSSION

Analysis based on disease, month and year of diagnosis for patients with malaria (Figure 3): This Visualizes some reports based on the disease, month, year and the total number of patients. This report provides the number of patients that suffered from malaria in each month, covering the period of 2020, 2021 and 2022.

Analysis based on patient's marital status and year of diagnosis for malaria (Figure

4): This shows the analysis of patient's record based on their marital status and year of diagnosis. This analysis is important for the decision maker to know the rate of disease infection among married and single patients of each year.

Analysis based on patient's age group and year of diagnosis for malaria (Figure 5): This shows data analysis for malaria cases between the age ranges of 25 to 30 years, the number of patients and the year of diagnosis. This analysis is important because it shows the decision makers the analysis of age group mostly affected by a particular disease, the year and the number of patients affected.

Data visualization based on disease, gender and year of diagnosis for malaria (Figure 6): This shows sample report based on disease, patient gender and the year of diagnosis. It tells the user the gender that is most hit by the disease within the period 2020, 2021 and 2022. It givesa detailed form of visualization of data for decision makers rather than the row and column form of record display.

Data visualization of disease case based on LGA of residence for Malaria (Figure 7): This shows sample report based on patients LGA of residence that were diagnosed with Malaria. It enables the user to know the LGA that is most hit by the disease within the period 2020, 2021 and 2022.

Data visualization of death rate based on marital status and year of diagnosis for malaria (Figure 8): This shows the death rate based on patient marital status and the year of diagnosis. It gives the user the death toll between married and single patients

within the periods 2020, 2021 and 2022 for Malaria disease.

Data visualization of death rate based on LGA of patients for Malaria (Figure 9): This shows the number of death recorded from Malaria based on their LGA of residence between the years 2020, 2021 and 2022. It enables the user to know the LGA that is most hit by the disease.

Analysis report based disease, month and year of diagnosis for Typhoid Fever (Figure 10): This shows some reports based on the disease, month, year and the total number of patients. The report tells the user the month and year the disease records its highest number of cases within 2020, 2021 and 2022.

Analysis report based on patient's marital status and year of diagnosis for Typhoid Fever (Figure 11): This displays the analysis of patient record based on their marital status and year of diagnosis. This is important for the decision maker to know the rate of the disease infection among married and single patients of each year.

Analysis report based on patient's age group and year of diagnosis for Typhoid Fever (Figure 12): This shows data analysis for Typhoid Fever cases between the age ranges of 20 to 25 years of age. This is important for decision maker(s) to know the age group that is mostly affected by a particular disease, as well as the year, and the total number of patients diagnosed within the period of 2020, 2021 and 2022.

Data visualization based on disease gender and year of diagnosis for Typhoid Fever (Figure 13): This shows sample report based on disease, patient gender and the year of diagnosis. This chart gives a

detailed form of visualization for the user to know at a glance, the gender with the highest number of disease cases within 2020, 2021 and 2022.

Data visualization based on patient's LGA of residence for Typhoid Fever (Figure 14): This displays a sample report based on patient's LGA of residence. It gives the user the total number of disease cases recorded by each LGA and the LGA with the highest cases of Typhoid Fever within the year 2020, 2021 and 2022.

Data visualization of death rate based on patient's marital status (Figure 15): This report gives analysis of death rate based on patient's marital status and the year of diagnosis. It helps the user to know at a glance, the death recorded among married and single patients. It also tells which group has the highest death rate within the periods 2020, 2021 and 2022 for Typhoid Fever case.

Data visualization of death rate based on patient's LGA of residence for Typhoid Fever (Figure 16): This reports the number of deaths recorded between the years 2020 to 2022 for patients that suffered from Typhoid Fever based on their LGA of residence. It allows the user to see at a glance, the deaths recorded in all LGAs and the LGA with the highest death rate.

CONCLUSION

This study designed and implemented a data warehouse for mining and simulating hospital records using data visualisation technique within the context of the healthcare service in other to better incorporate patient records into single systems for simpler and improved data mining, analysis, reporting and querying.

The data warehouse contains only data that is required for data mining, reporting and analysis for the purpose of this study and it can be updated periodically, such that all the data can be integrated from different sources into the central data warehouse

The study demonstrated how data can be incorporated from diverse desperate heterogeneous clinical data stores into a single data warehouse for mining and analysis purpose to aid medical practitioners and decision makers in decision making.

The Framework for data warehouse and Mining of Hospital record of patients was designed not to be specific to only a number of diseases but to accept as many diseases as possible without any limitation.

The designed framework can be used by industry professionals and researchers for implementing data warehousing system in the medical field. It can also be used for community diagnosing in cases of outbreak of certain disease .Developing a Framework for data warehouse and Mining of hospital records is very essential, particularly for medical decision-makers, academic researchers, IT professionals and non-professional. Clinical data mining must not only support medical professionals and decision makers to understand the past but also it strives professionals to work towards new prospects.

REFERENCES

Abubakar A., Aliyu, A., Bello, S. A., & Gezawa, A. S. (2014). Building a Diabetes Data Warehouse to Support Decision making in healthcare industry, 16(2), 138–143.

Adeleke, I. T., Lawal A. H., Adeleke. R. A. & Abubakar. A. A. (2014). Health information technology in Nigeria: Stakeholders' Perspectives of nationwide implementations and meaningful use of the emerging technology in the most populous black nation. *American Journal of Health ResearchSpecial Issue: Health*

Information Technology in Developing Nations: Challenges and Prospects Health Information Technology, 3(1–1), 17–24.

Ahmadu, A. S., Boukari, S., Garba, E. J., & Gital, A. Y. (2017). Visualisation of students' Academic Performance using Human Learning System. *International Journal of Scientific & Engineering Research, 8*(10).

Arunachalam, P. a. (2017). Healthcare data warehousing. *i-manager's Journal on Computer Science, 4*, 2-6.

Avati A., Jung K., Harman S., Downing L., N. A. & S. N. . (2017). Improving palliative care with deep learning. *Journal of Medicine Internet Research, 7*(2), 1318.

Başaran, & Beril, P. (2005). *A Comparison of Data Warehouse Design Models.* Atilim University, The Graduate School of Natural and Applied Sciences, Turkey.

Bateman N.D, C. A. . & G. K. . (2010). An audit of the quality of operation notes in an otolaryngology unit. *Journal of Research College & Surgery Edinburg, 44*(2), 92–95.

Bum Chul Kwon, Min-Je Choi, Joanne Taery Kim, Edward Choi, Young Bin Kim, Soonwook Kwon, Jimeng Sun, and Jaegul Choo (2019). RetainVis: Visual Analytics with Interpretable and Interactive Recurrent Neural Networks on Electronic Medical Records. Citation information: DOI 10.1109/TVCG.2018.2865027, IEEE Transactions on Visualization and Computer Graphics

Demarest, M. (2014). *Data Warehouse Prototyping: Reducing Risk, Securing Commitment and Improving Project Governance.* Retrieved from wherescape.com: http// www.wherescape.com.

Desouza, K. C., & N. Wickramasinghe, J. N. . G. and S. . S. (2015). Knowledge management in hospitals. *Creating Knowledge Based Healthcare Organizations*, 14–28.

Epstein, I. (2010). *Clinical data-mining: Integrating practice and research.* London: Oxford University Press.

Faramarz P, Hossein M, Alireza K, Johan E, U. F. (2018). What they fill in today, may not be useful tomorrow: lessons learned from studying Medical Records at the Women hospital in Tabriz, Iran. *BMC Public Health, 8*(3), 110–139.

Frank, L., & Andersen, S. (2010). Evaluation of different database designs for integration of heterogeneous distributed electronic health records. *The 2010 IEEE/ICME International Conference on Complex Medical Engineering*, (204 – 209).

Güzin, T. (2017). Developing a Data Warehouse for a University Decision Support System. *The Graduate School of Natural and Applied Sciences, Atilim University, 6*(3), 714.

Huang, C., Lu, R., Iqbal, U., Lin, S., Anh, P., Nguyen, A., Yang, H., Wang, C., Li, J., Ma, K., Li, Y. J., & Jian, W. (2015). A richly interactive exploratory data analysis and visualization tool using electronic medical records. *BMC Medical Informatics and Decision Making*, 1–14. https://doi.org/10.1186/s12911-015-0218-7

Jothi, N., Aini, N., Rashid, A., & Husain, W. (2015). Data Mining in Healthcare – A Review. *Procedia - Procedia Computer Science, 72*, 306–313. https://doi.org/10.1016/j.procs.2015.12.145

Kimball,D. K. & Ross, M. R. (2002). The Data Warehouse Toolkit: The Complete Guide to Dimensional Modelling. In R.

Kimball, & M. Ross, *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modelling.* New York: John Wiley & Sons, Inc.

Mishra., D., Kumar, A. D., Mausumi, & Mishra., S. (2010). Predictive Data Mining: Promising Future and Applications. *International Journal of Computer and Communication Technology, 2*.

Saliya, N. (2013). *Data warehousing model for integrating fragmented electronic health records from disparate and heterogeneous clinical data stores.* Queensland University of Technology, School of Electrical Engineering and Computer Science Faculty of Science and Engineering, Australia.

Swinton, P. A., Hemingway, B. S., Saunders, B., & Gualano, B. (2018). *A Statistical Framework to Interpret Individual Response to Intervention : Paving the Way for Personalized Nutrition and Exercise Prescription.* *5*(May). https://doi.org/10.3389/fnut.2018.00041

Vankipuram, A., Traub, S., & Patel, V. L. (2018). A method for the analysis and visualization of clinical workflow in dynamic environments. *Journal of Biomedical Informatics*, *79*, 20–31. https://doi.org/10.1016/j.jbi.2018.01.007

Wager, K; Lee, F. and Slaser, I. P. (2015). Managing health care information system. *Journal of Healthcare Engineering*, *6*(4), 10–20.