A CENTRALIZED DATA WAREHOUSE APPROACH FOR HEALTH INFORMATICS ENVIRONMENT

[1]R.Abiya Rajarathinam, [2]R.Augusta Thangam, [3]A.Divya, [4]P.Jenifer Mary, [5]A.Jenushiya
Assistant Professor, Dept of IT, Kings Engineering College,
[2,3,4,5]UG Students, Dept of IT, Kings Engineering College.

ABSTRACT

The aim of this paper is to propose a Centralized data warehouse approach for health informatics environment. An existing healthcare industry clinicians and researchers are using multiple data sources to obtain the data for their research and other quality improvement purposes. There is no single database or infrastructure exists in current healthcare industry. Our proposed data warehouse can be used directly by outcome researchers and clinical investigators to facilitate data access for the purposes of hypothesis testing, cohort identification, data mining, and risk prediction. Our proposed approach could reduce the health care cost and enhanced the treatments for disease, increase the efficiency of new medical clinical informatics environment and facilitate sharing and reuse of patient data. The twin pressures of cost containment in the healthcare market and new federal regulations and policies have led to the prioritization of the meaningful use of electronic health records in the United States. EDW and SIA layer on top of EDW is becoming an essential strategic tool to healthcare institutions and integrated delivery networks in order to support evidence based medicine at the enterprise level.

***Index Terms*-clinical Data Warehouse, Smartphone Health App, Cohort Identification, Natural Language Processing, Readmission Risk, Outcomes Research.**

I. INTRODUCTION

The clinicians and researchers at Houston Methodist Hospital (HMH) system use multiple data sources to acquire the data for research and quality improvement purposes as no single infrastructure or database exists that could provide them with ease the data required for their research[18]. HMH system is home to seven hospitals and operates approximately nine major categories of clinical databases. Current methods of obtaining data from all these locations and vendors for preparatory-to-research questions often involves laborious, time-consuming manual extracts and cleansing of data for specific projects. It is recognized that the current process is cumbersome, costly and time consuming and adds no intrinsic value to the research being undertaken. This leads investigators to spend a lot of unproductive [2] time in negotiating and waiting for data instead of conducting the research. Worse, the data ultimately delivered often are incomplete, depending on the understanding and knowledge of the person retrieving the data. In many institutions, a "gray market" for data could develop, as researchers find unofficial workarounds to obtain data they need for their work. This “gray market” approach could lead to compliance and security risks, as isolated silos of patient data evolve in different parts of the healthcare organization without formal oversight for [9]Health Insurance Portability and Accountability Act (HIPAA) and Institutional Review Board (IRB) compliance, and outside of the processes for protecting data from misuse or breach. HMH researchers need access to vast pools of patient data to develop and test their[7] scientific hypotheses, so the making of a solitary, integrated data system would provide a huge opportunity for an expanded number of biomedical research projects, including large-scale projects.

t provides access to users to build queries to make a clinical profile and construct a study cohort based on EMR[15] data by stringing series of search elements together. Nevertheless, i2b2 has its own limitations [4]. It cannot provide any data beyond the number of patients matching the search criteria and is incapable of running time-based queries, e.g., number of admissions in a month or events that happen before and after a clinical event or a laboratory result within a certain time period. It is also incapable of tasks such as sample tracking, clinical trials, study administration, and real-time alerts. Undoubtedly, even with the most recent integration of natural language processing into i2b2 [5], it is difficult to imagine that the complete case control building process could be finished consequently without critical restructuring of clinical data inputs format.

Centralized data warehouse was designed and developed as a platform that allows HMH system to im-

prove rather than just to report on performance and to provide a reliable and coherent platform for developing a broad range of novel clinical solutions at the software intelligence and analytics (SIA) layer to support the practice of evidence-based medicine at the enterprise level. Centralized data warehouse data warehouse provides data access that contains records dating back to January 1, 2006, with over 2 million unique patients and over 10 million unique patient encounters that encompass a range of diverse, heterogeneous data, e.g., demographics, pathologies, imaging, laboratory tests, physical examinations, and treatment regimens. Centralized data warehousecould inevitably prompt decreased health care costs and enhanced treatments for disease, increase the efficacy of new medicaltechnologies, and eventually enhance patient satisfaction.

II. MATERIALS AND METHODS

As data is collected from all the locations of HMH system for centralized data warehousing, the aim of Centralized data warehouse is to provide a uniform, secured yet flexible[11] information environment for supporting clinical and outcomes research by integrating, categorizing, and analyzing vast amounts of clinical, demographic, genomics, imaging, administrative, and financial data from the entire hospital system as well as external and internal research databases . The framework of centralized data warehouseconsists of two components: the enterprise data warehouse and a [8]software intelligence and analytics (SIA) layer for enabling a wide range of clinical decision support (CDS) system.

ARCHITECTURE DIAGRAM:

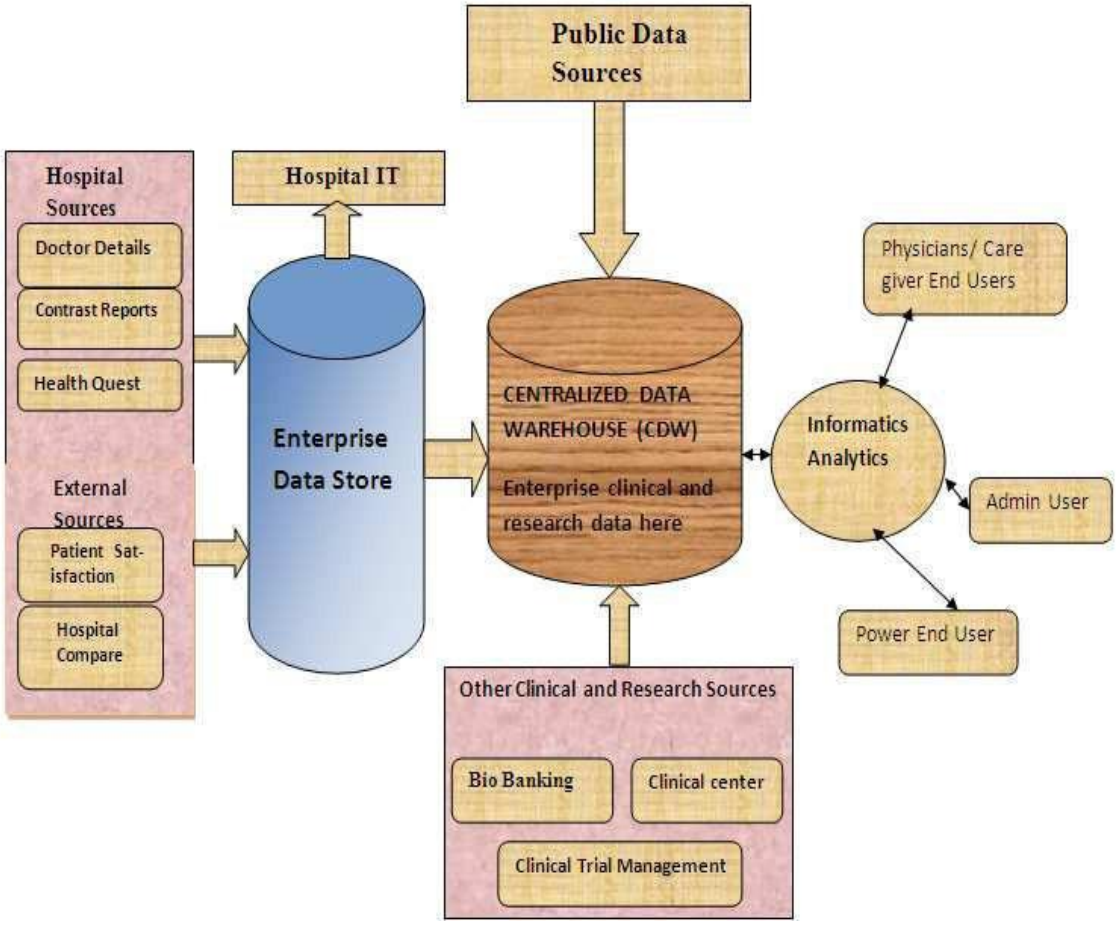

Fig 1.Overview of centralized data warehouse structure

A. Data Warehouse

The back-end database of centralized data warehouse expands on the i2b2[9] S-schema with different visual analytics, presentation logic, and query interfaces created for different groups of users, e.g., physicians and caregivers, hospital administrators and executives, power users such as biostatisticians[12], and automated report generation for outcomes and clinical research purposes. The clinical informatics team solicited user requirements from physicians such that the user requirements drive functional requirements for implementation and evaluation. The physicians and the academic[3] investigators require vast pools of patient information to advancing science and/or medical practice. An individual patient record usually consists of multiple binary (e.g. gender), categorical (e.g. race) and continuous (e.g. age) attributes at a given time point, time-lapse information (e.g. multiple blood pressure reads during a patient's stay in the hospital), and non-alphanumeric data (e.g., free text and images). Like other scientists, clinical investigators participate in thorough review work before continuing with a research project that may take years to finish.

The development of[14] Centralized data warehouse data warehouse required several steps. Initial activities included seeking technical and administrative support, concluding which data elements were desired, and learning how and where the data were stored in each hospital's information systems. Each clinical department (e.g., medical records, laboratory, and pharmacy) at HMH stored data in a separate database server. Obtaining information about each database's architecture, such as [13]data dictionary and documentation about the database tables' relationships, was essential for understanding the data and for extracting relevant variables into Centralized data warehouse enterprise data warehouse. Centralized data warehouse data warehouse continues to be developed and implemented in phases with the close collaboration with [15]Houston Methodist Hospital IT. Hospital IT maintains the data repository and its software applications in each phase of production while theCentralized data warehouse team continues to work on the next phase of development.

B. CDS Application at the Software Intelligent ant Analytics(SIA)Layers

CDS tools are the front-end informatics and analytics tools that include web and desktop access, smart phone interface, and[16] sensor monitoring. CDS systems include information retrieval, alerting systems, suggestion systems, and prediction models [5]. centralized data warehouse allows effective integration and management of multiple data types to apply biostatistics and machine learning for evidence-based decision supports. In this paper, we present three clinical solution applications in the SIA layer of the centralized data warehouse framework, namely, we: (1) established and implemented mPOD (CENTRALIZED DATA WAREHOUSE Point of Delivery), a real-time interaction tool that is used to query and collect population health data and provide clinical.

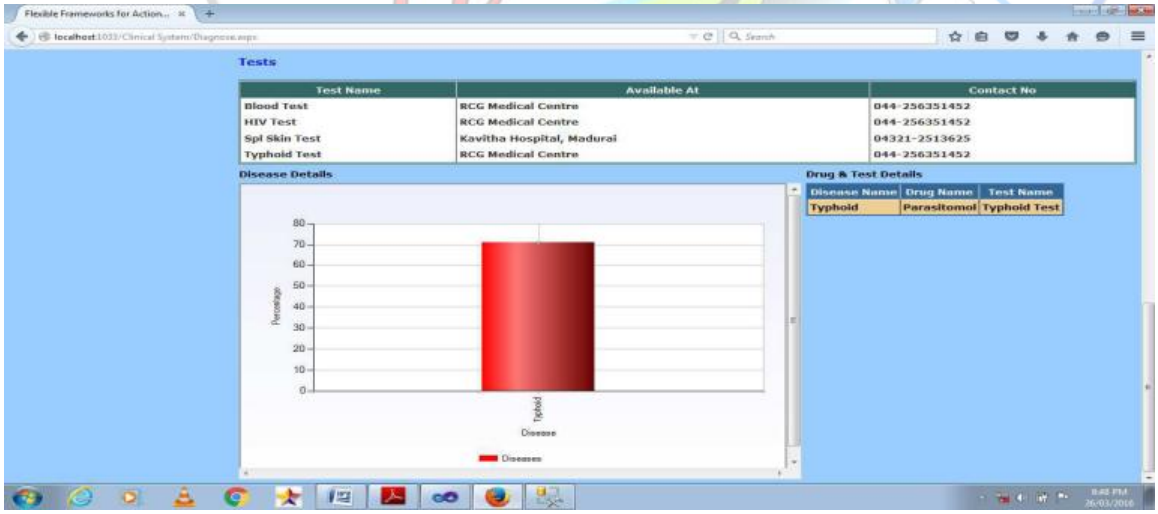

Fig2 .Percentages oftyphoid patients - Centralized data warehouse

- **Category 1: Information Retrieval**
  Data mining is meant to extract the hidden knowledge in the data and to reduce the overall cost of research and health care. The objective of the first application is for mPOD (CENTRALIZED DATA WAREHOUSE Point of Delivery) to be directly used by clinical investigators, for the purpose of data exploration to simplify the process of using existing clinical data for potential hypothesis and preliminary research cohort discovery. A second application is the application of natural language processing techniques that will allow researchers to efficiently search through patients' medical reports, laboratory results, and any other records. The free text reports are manually written by the physicians and the clinical information can be extracted by reading these reports. Due to the big data clinical reports in the centralized data warehouse, it is a challenge to screen and extract all patients' clinical information from the large number of reports without automated methods. Therefore, we developed a Natural Language Processing (NLP) module to assist in extracting this information from free text reports.
- **Category 2: Prediction Models**
  Prediction models are the most evident and straightforward targets for applying data mining algorithms. We illustrated this with another CDS application, READMIT, which is used to analyze both retrospective and real time clinical data in centralized data warehouse to stratify readmission risk of patients. Most of these readmissions seem attributable to medical issues different from those responsible for the initial admission. READMIT is developed to predict a patient's readmission rate based on their medical record using support vector machine (SVM) methods.

Fig3. Details about patient history from Centralized data warehouse

1) mPOD(Clinical Data Ware House Point of Delivery)

Tools for querying data have usually been text based even though graphical interfaces [14]have been pursued . mPOD is the web enabled query interface being developed on top of the centralized data warehouse EDW to enable investigators and physicians to quickly search for entire patient populations at Houston Methodist Hospital system and identify patient cohorts for clinical trials or outcome studies. In order to support hypothesis generation and testing andcohort identification, the key challenge is an interface that greatly accelerates access to relevant data sets: past queries should be quickly recallable; new queries should be easily constructible; existing queries should be readily modifiable. The mPOD application

includes features to support HIPAA compliance, full and privileges, integration with the institutional LDAP server, and a dashboard displaying population statistics of data . The row of criteria button items is the "vocabulary" of the database. Even though the underlying structure of the database places the items in completely different tables and structures, to the user everything appears to be just another item to be selected from the criteria list. Once the criteria are selected, the user is able to add constraints for the items. If an item is associated with a value, like a laboratory test, then a selection could be made to define the value. The value allows for "No value", meaning all tests are considered valid, a flag selection option from the performing lab to identify "High/Low/Abnormal/normal" value, or an inequality number, which allows the user to specify tests greater than, less than or in between specified values. When the query is run, aggregate numbers on the patient population matching the query criteria are returned in the results section of the page. Typically, results are limited to counts and aggregate statistics until the user achieves a sense of which direction to pursue further. We obfuscate the results somewhat to prevent any possibility that a single individual may be identified through the interface. Once a set of patients is obtained with the[17] Query tool and the user is satisfied they match the general characteristics required for their research study, the individual patients may be identified, provided the Institutional Review Board has approved the study. The identified list of patients is delivered to the researcher in the format required by the researchers that include all electronic data for those patients.

To support the simple visual model, as well as the identification and distribution of patient data, a three-tier software architecture[18] was employed. The user interface was developed using Java Server Faces, the middle layer was developed in Java on top of spring framework, and a Microsoft SQL 2008 database is queried and managed in standard SQL.

2) MOTTE (Methodist hospital Text Teaser)

Extraction of cancer-related parameters from pathology reports is essential for many clinical studies, however researchers and cancer registrars generally depend on manual extraction of information, a time intensive and costly process. However, no reliable commercial software exists with customized workflow and application logic to fill our needs. Thus, we developed a [7]natural language processing (NLP)[16] based tool MOTTE to extract and structure data in pathology reports automatically to support our clinical solution applications. Such a tool can search and retrieve specific clinical information from audit trails, access control free text reports archived in centralized data warehouse enterprise data warehouse.

Pathology reports were collected associated with breast biopsies that were referred for evaluation of cancer prognosis. Normally, different clinical information belongs to specific sections of the report. Hence, we first determine the general structure of the reports and recognize the data required from each section of the report. Then, MOTTE was applied to search and retrieve specific clinical parameters from specific section of the report.

The NLP based method contains five processing steps, i.e.; tokenization, stemming, stop words removal, vector space modeling, and similarity calculation . The first step involves tokenization that translates to turning each clinical text report into a stream of tokens.

3) READMIT (Readmission Risk Estimator)

The goal of READMIT application is to establish the relationship between readmission risks of the patient with the medical information[17] of patient's record after initial admission. This optimization actually corresponds to a ridge regression in feature space. The Lagrangian method is utilized to solve the constrained optimization problem. We implemented the model in the READMIT application tool so that the real time clinical data in centralized data warehouse then can be used for stratifying readmission risk of patients.

III. RESULTS

Usability analysis was performed as a preliminary evaluation of centralized data warehouse components. Listed below are examples of the patient outcomes research studies supported by the centralized data warehouseEDW[3]. Another common use of the EDW[19] is to support hospital quality improvement efforts.

A. mPOD (CENTRALIZED DATA WAREHOUSE Point Of Delivery)

The mPOD (CENTRALIZED DATA WAREHOUSE Point of Delivery), enables investigators and physicians to quickly search for entire patient populations at Houston Methodist Hospital system and identify patient cohorts for[5]clinical trials or outcome studies. In order to understand if the query tool could indeed be used successfully by novice users, a usability analysis of the tool was performed. All queries submitted to the database are saved and thus can be anal[10]yzed. Upon analyzing the thirty selected re-

search data requests, it was observed that about 43% ofthese data requests could be completed using the mPOD alone, without demanding any significant changes to the ontology. About 67% of the data requests needed actual patient data in addition to the counts. Frequently requested data elements were demographics, visit details (visit age, visit type or length of stay), procedures, diagnoses, medications, and laboratory tests. An example of a data request is provided here. In a proposed erectile dysfunction clinical study, the following inclusion and exclusion criteria were queried in mPOD to yield a cohort of patients that met the strict set of requirements to participate in the study. Inclusion criteria were Male; Age 35 – 65 years; diabetes (ICD 9 code: 250.0*); A1C > 6.5, BMI > 30 and prior diagnosis of organic impotence (ICD 9 code: 607.84). The criteria that excluded patients from being part of the cohort include history of any of the following: Coronary Artery Disease including Myocardial Infarction (ICD 9 codes: 401.0 – 405.99); Cardiac arrhythmias (ICD 9 codes: 427.0–427.9); and current use of any of the following medications: Propanolol, Sildenafil citrate, Vardenafil, Tadalafil. All the aforementioned data requests were fulfilled using the mPOD alone.

Amongst the data requests that could not be fulfilled with the mPOD without major modifications, eight cases had one or more precise criteria (for example, data from the outpatients database, specific service lines, etc.), and some had one or more temporal criteria that could not be created using the mPOD alone. Such query criteria often required pre- or post-processing of data to answer the research questions. During the sample data requests analysis, it was also discovered that 48% of these requests required some type of pre-processing of the data, 59% required some post-processing, and 29% of the requests required both the processing in order to fulfill the data request. Some requests included calculated fields, and other requests required additional attributes in order to accommodate one or more complex inclusion and exclusion criteria.

B. MOTTE (Methodist hOspital Text Teaser)

To validate the tool, 670 breast cancer candidate patients' biopsy reports were used in the application. For the data extraction, the [10]physicians provided a list of key words from the clinical information required to extract cancer patients. All items from the key words list were used to compute the similarity measure with the sentences. By comparing the results using our method and physicians' manual method, a threshold 0.4 for similarity measure was set as an optimal value; if the maximal Jaccard similarity coefficient was greater than 0.4, then we determined that the item from clinical information table was positive in the text report. Moreover, the threshold 0.4 is the optimal value for all reports in our experience, the users normally do not need to change the threshold when using MOTTE for other types of reports.

We implemented our framework in Microsoft Visual Studio 2010 environment. All of the programs were performed on an Intel workstation with Quad Core 3.5 GHz CPU and 16GB memory running Windows 7. We used Java to program our method and developed a web interface for the application. The time consumption was three hours for retrieving all information. To ensure effectiveness, the results of MOTTE were monitored by the physicians and compared with the physicians' manual results.All 670 breast cancer candidate patients grouped by MOTTEWith our method, we found 33 cases to be benign, 4 patients to be Atypical, 0 cases to be LCIS, 55 cases to be DCIS, and 503 cases to be Carcinoma. Accurately and promptly figuring out carcinoma patients is the main purpose for the MOTTE utilization, so the physicians helped us to find the carcinoma patients from the biopsy reports manually as validation. Using the manual method, the percentage of the carcinoma patients from all cancer candidate patients is 77.9%, and the percentage using MOTTE is 75.1%. Overall, an acceptable range of accuracy was obtained for extracting the carcinoma patients using MOTTE. Armed with these results, we also help the physicians to complete the data analysis for assisting the cancer risk assessment. From the results, the physicians can conclude the relationship between the mammogram density and risk of breast cancer.

C. READMIT (Readmission Risk Estimator)

The READMIT application is a smart phone app that uses SVM method and learns from patient information stored in the clinical data warehouse or electronic medical record [20]. The medical information for training SVM model include (1) length of stay in hospital, (2) number of co-morbid diagnosis present on admission, (3) emergency as portal for admission, and (4) the severity of admit illness. These four variables were identified that were significant ($p<0.05$) in predicting readmission.The users need to fill the four major features into the application, and click on the "check risk" button. Based on the values provided, the application calculates a percentage. If it is below 0.5, it's low risk, otherwise, high risk of readmission. Currently the accuracy of readmission prediction of the 3-fold cross validation is about 70.9%. According to the feedback of physicians, this accuracy is acceptable. READMIT app is now under satisfaction evaluation and will be published in the App stores with secured password access. Based on the values provided, the application calculates a percen-

tage. If it is below 0.5, it's low risk, otherwise, high risk of readmission.

IV. DISCUSSION

Clinical and translational[20] research requires integration of clinic pathologic, bio-specimen, phenotypic and molecular data, across multiple data platforms. It is also critical to make these data usable whenever required. Enterprise data warehousing, as an important component of health informatics, provides a powerful means for data integration and efficient access in supporting clinical and translationa[9]l research. Centralized data warehouse takes a top-down "Enterprise" approach with open specialized functions and applications . It addresses significant niches in the clinical informatics area, for example, providing researchers and clinicians with sufficient access to clinical data to enable study design and cohort selection, while reducing many of the common concerns and patient privacy risks.

The web query interface is frequently recognized as a bottleneck to evaluate the efficiency of different investments and revolutions in clinical research [13]. User-centered design process [12] and agile development methodologies are used for the development of query interface. Currently, very few interfaces are in use directly by the investigators. Features of query interface design that facilitate its use by investigators include efficient query construction, obtaining counts in real time, query sharing and reuse, and data exploration. The goal of mPOD is to develop a powerful interface that is simple, usable, and intuitive and also to serve as an on-line informatics[11] tool by investigators and physicians to facilitate data access for the purposes of cohort identification, screening, data mining, clinical research training, case-control comparison and longitudinal studies.

Creating and implementing a readmissions reduction strategy should be a top priority for any hospital that isn't measuring up in this area. A successful strategy for reducing readmissions will involve identifying at-risk patients for readmission[14], bridging the gap in transition care planning, tracking at-risk patients with device-enabled real time notification of care providers, and monitoring performance improvements as a result of these efforts. The right clinical analytics application can help hospitals harvest the right data, turn these data to a source of wisdom to transform clinical care workflows, and measure the precise quality and impacts of all of these activities.

V. CONCLUSION

The enhancement of the hospitals usage by the patients to select a best doctor and best hospital from the location by quick search option. The search option may done by the keyword search or hospital name search or doctor name search or by the specialty search .By searching this way patient can easily select the best doctor for their treatment for that the patient have to book the appointment for the specific specialty doctor for a feasible date. The doctor can view his list of patients appointment. If the doctor may not be available on that particular date means ,the doctor can able to cancel and reschedule the particular patient appointment .After that if the patient satisfied for that doctor reschedule process then the patient can take that appointment. This is the process of this applications. This application will help to reduce theusers difficulty of doctor selection.

REFERENCES

[1] Murphy, Shawn N. et al. (2007). Architecture of the Open-Source Clinical Research Chart from Informatics for Integrating Biology and the Bedside.AMIA Annual Symposium Proceedings. pp. 548–552.

[2] Weber, Griffin M. et al. (2009). The Shared Health Research Information Network (SHRINE): A Prototype Federated Query Tool for Clinical Data Repositories. Journal of the American Medical Informatics Association. 16(5), pp.624–630.

[3] Murphy, Shawn N. et al. (2006). Integration of Clinical and Genetic Data in the i2b2 Architecture.AMIA Annual Symposium Proceedings.2006, p. 1040.

[4] Johnson, Emilie K et al. (2014). Use of the i2b2 Research Query Tool to Conduct a Matched Case–control Clinical Research Study: Advantages, Disadvantages and Methodological Considerations. BMC Medical Research Methodology. 14(1), p. 16.

[5] Heinze, Daniel T. et al. (2008). Medical i2b2 NLP Smoking Challenge: The A-Life System Architecture and Methodology. Journal of the American Medical Informatics Association. 15, pp. 40–43.

[6] Murphy, Shawn N et al. (2010). Serving the Enterprise and beyond with Informatics for Integrating Biology and the Bedside (i2b2). *Journal of the American Medical Informatics Association.* 17(2), pp. 124–130.

[7] Sahama, Tony R., and Peter R. Croll. (2007). A data warehouse architecture for clinical data warehousing. *Proceedings of the fifth Australasian symposium on ACSW frontiers. 68*, pp. 227-232.

[8] Ledbetter, Craig S., and Matthew W. Morgan. (2001). Toward best practice: leveraging the electronic patient record as a clinical data warehouse. *Journal of healthcare information management.* 15(2), pp. 119-132.

[9] Kennedy, G.D. (2013). Enterprise Informatics: Key to Precision Medicine, Scientific Breakthroughs, and Competitive Advantage. *Remedy Informatics*.

[10] Zerhouni, Elias A. (2005). US biomedical research: basic, translational, and clinical sciences. *Jama.* 294(11), pp. 1352-1358.

[11] Zerhouni, Elias A. (2006). Clinical research at a crossroads: the NIH roadmap. *Journal of Investigative Medicine*. 54(4), pp. 171-173.
[12] Wong, Stephen T.C. et al. (2002). Design and Applications of a Multimodality Image Data Warehouse Framework. *Journal of the American Medical Informatics Association. JAMIA*. 9(3), pp. 239–254.
[13] Calinescu, Radu, et al. (2007). Cross-trial query system for cancer clinical trials." *Innovations and Advanced Techniques in Computer and Information Sciences and Engineering*. pp. 385-390.
[14] Highsmith, et al. (2001). Agile software development: The business of innovation. *Computer*. 34(9), pp. 120-127.
[15] Berner, Eta S. (2009, June). Clinical decision support systems: state of the art. *AHRQ Publication.* 09-0069-EF, pp. 4-26.
[16] Nigrin, Daniel J., and Isaac S. Kohane. (1998). Data mining by clinicians. *Proceedings of the AMIA Symposium*. pp. 957-61.
[17] Banhart, F., and H. Klaeren. (1995). A graphical query generator for clinical research databases. *Methods of information in medicine.* 34(4), pp. 328-339.
[18] Bales, Donald. (2002). *Java programming with Oracle JDBC.* Available: www.java.com
[19] *Handbook of Natural Language Processing*, 2nd ed., Computational Linguistics, 2011, 37(2), pp. 395-397.
[20] Abbasi, A., et al. (2014). Social Media Analytics for Smart Health. *IEEE Intelligent Systems.* 29(2), pp. 64-67.