3-2022

A Privacy-Preserving National Clinical Data Warehouse: Architecture and Analysis

Md Raihan Mia
Bangladesh University of Engineering and Technology

Abu Sayed Md Latiful Hoque
Bangladesh University of Engineering and Technology

Shahidul Islam Khan
International Islamic University Chittagong (IIUC)

Sheikh Iqbal Ahamed
Marquette University, sheikh.ahamed@marquette.edu

Smart Health, Vol. 23 (March 2022): 100238. DOI.

A Privacy-preserving National Clinical Data Warehouse: Architecture and Analysis

Md Raihan Mia
Computer Science and Engineering Department, Bangladesh University of Engineering and Technology, Bangladesh
Abu Sayed Md Latiful Hoque
Computer Science and Engineering Department, Bangladesh University of Engineering and Technology, Bangladesh
Shahidul Islam Khan
Computer Science and Engineering Department, IIUC, Bangladesh
Sheikh Iqbal Ahamed
Computer Science Department, Marquette University, USA

Abstract

A centralized clinical data repository is essential for inspecting patients' medical history, disease analysis, population-wide disease research, treatment decision support, and improving existing healthcare policies and services. Bangladesh, a rapidly developing country, poses several unusual challenges for developing such a

centralized clinical data repository as the existing Electronic Health Records (EHR) are stored in unconnected, heterogeneous sources with no unique patient identifier and consistency. Data integration with secure record linkage, privacy preservation, quality control, and data standardization are the main challenges for developing a consistent and interoperable centralized clinical data repository. Based on the findings from our previous researches, we have designed an anonymous National Clinical Data Warehouse (NCDW) framework to reinforce research and analysis. The architecture of NCDW is divided into five stages to overcome the challenges: (1) Wrapper-based anonymous data acquisition; (2) Data loading and staging; (3) Transformation, standardization, and uploading to the data warehouse; (4) Management and monitoring; (5) Data Mart design, OLAP server, data mining, and applications. A prototype of NCDW has been developed with a complete pipeline from data collection to analytics by integrating three data sources. The proposed NCDW model facilitates regional and national decision support, intelligent disease analysis, knowledge discovery, and data-driven research. We have inspected the analytical efficacy of the framework by qualitative evaluation of the national decision support from two derived disease data marts. The experimental result based on the analysis is satisfactory to extend the NCDW on a large scale.

Keywords

Privacy, Standardization, Clinical data warehousing, Data modeling, Big data analytics, Decision supports

1. Introduction

The velocity of Electronic Health Records (EHR) generation has increased exponentially over the past decade, with growing data repositories across healthcare systems (Evans, 2016). The leverage of the rising Big Data to create large datasets for data mining, artificial intelligence, machine learning, deep learning, and research required data integration, record linkage, quality control, storage, retrieval, and exploitation. There are numerous current research areas within the field of Clinical Informatics. For instance, intelligent decision support systems for meaningful use of EHR (Elliott et al., 2021, Musen et al., 2021), combating COVID-19 (Sherimon et al., 2021), AI enabled expert system (Wang et al., 2021), etc.; micro-level disease analysis like chronic kidney diseases (Hagar et al., 2014), predictive analysis of diseases (Yang et al., 2021); public health surveillance, trend analysis, and prevention (Birkhead et al., 2015, Calman et al., 2012, Perlman, 2021); and also personalized patient care (Devi & Rizvi, 2022).

As a fast-growing developing country, analysts and researchers of Bangladesh cannot utilize clinical data because of the discrete nature of current and historical clinical records in different operational systems without record linkage. In Bangladesh and other economically developing countries, people do not have medical cards with unique health ID and healthcare centers do not store National ID numbers or Social Security Numbers (SSN) which leads to a lack of suitable datasets for research and analysis (Khan & Hoque, 2017). Consequently, inadequate disease analysis, health risk computation, forecasting, decision-making, research, etc., leading Bangladesh to deteriorate healthcare services, severe health crises, and mal-distribution in national and regional health coverage (Al-Zaman, 2020, Joarder et al., 2019, WHO, 2021). A report on Harvard Edu pointed to an emerging global health crisis of non-communicable diseases, especially in Low- and Middle-Income countries because of the insufficiency of the national health and regulatory systems (Daniels & Donilon, 2017).

A centralized clinical data warehouse can mitigate the above problems by providing a ubiquitous research data platform and data-driven analytical solutions. Developing a clinical data warehouse is resource-intensive and time-consuming but essential for decision support, knowledge discovery, and research to deliver quality health services. The main challenges for developing a clinical data warehouse from the noisy health data of Bangladesh are including:

1. Data privacy and security to ensure patient safety.

2. Data integration from heterogeneous sources with secure record linkage.
3. Data standardization, data exchange and interoperability framework development.
4. Developing a highly secure and scalable clinical big data repository in a flexible, query-able format.

In the department of CSE, BUET, several research studies have been conducted to resolve the potential underlying problems to develop an integrated health data warehouse from noisy data (Khan & Hoque, 2020a). Shahidul et al. presented a discussion on the existing health data management systems and the available features and limitations of the National Health Data Warehouse of Bangladesh in Khan, Hoque, and Ullah (2016). In Khan and Hoque (2016a), they provided essential recommendations to boost the integration process with supporting record linkage based on analyzing the practical problems of collecting and integrating healthcare data in Bangladesh to build up a central national health data warehouse. They have identified the prospects and complexities of health data warehousing in Bangladesh and proposed a suitable data warehousing model for integrating data from different healthcare sources (Khan & Hoque, 2015). Health data integration with secure record linkage is essential for privacy-preservation and knowledge discovery. Key-based Secured Record Linkage (KSRL) algorithm that can anonymize patients' identification while maintaining secure record linkage in the national health data warehouse has been developed with 96.42% accuracy, 93.3% precision, 96.4% recall, and 94.8% F-measure (Khan & Hoque, 2019). Webber et al. presented the architecture of Smart Health Center for integration and access of patient-centric transnational medical records that preserved privacy for data analytics (Webber, Santana, Vermeulen, & Bowles, 2020). Effective data mining strategies, such as Temporal impact analysis, Impact analysis of diagnostic tests, Fraud Testing Awareness, etc., from the clinical data warehouse, have been illustrated in Khan and Hoque, 2016b, Khan et al., 2015.

In the security context, digital health data containing protected health information (PHI) are the main target of cybercriminals. Humphrey et al. recommended the two most critical concerns for the up-gradation of Health Insurance Portability and Accountability Act (HIPAA) regulation to ensure increased awareness of the use of AI in healthcare and protect patient data (Humphrey, 2021). In Khan, Hoque, et al. (2016c), a state of the art review has been provided for the security threats in the integrated healthcare information systems and recommended some important techniques to minimize the risk of attacks and reduce the chance of compromising patient privacy after any successful attack.

In this research, a novel framework of the National Clinical Data Warehouse (NCDW) of Bangladesh has been developed by integrating three operational databases of laboratory diagnostic systems; that synthesized our previous research findings. The data integration has required unique technical, semantic, and ethical challenges (Alhazmi, 2019). However, the ethical clearance from Bangladesh Medical Research Council (BMRC) to collect anonymous secondary data (Ref: BMRC/NREC/20I9-2022/342) and presented techniques helped overcome these challenges. Data standardization and quality control methods enabled us to merge heterogeneous data in a consistent storage format. The multidimensional data model of NCDW materialized to process big data in an optimized and efficient way. NCDW facilitates regional and national decision support, intelligent disease analysis and management, powering knowledge discovery at the patient and population level. Moreover, the fusion of clinical data in a central repository will reinforce data-driven research at the micro-level.

Our contributions to face these unique challenges add to clinical informatics research are as follows:

1. We present a framework of the National Clinical Data Warehouse of Bangladesh. We design our data platform with the consideration of merging heterogeneous clinical data from dispersed sources. It is an effective framework to leverage large clinical research datasets, including pathology, radiology,

genomics, molecular, behavior, ambient data. This model can be used in similar challenged countries including population-wise large countries e.g., India.

2. We present the patient safety, data standardization, quality control, and data management techniques. We propose a wrapper-based data integration process that is effective for data anonymization, transformation, and uploading consistent data into the NCDW considering the data noise and disparity. Then, we collect real data from hospitals and analyze the ICD-11 diseases from the perspective of national decision supports.
3. Our experimental results show that our framework can reinforce data-driven research and analytical solutions. The proposed methodology is highly effective for developing countries where the national health standards and policy are absent for the clinical data structures and interoperability.

This paper is organized as follows. We have presented selected literature reviews on existing clinical data warehouses, design methodologies, and usability of Clinical Data Warehouse in Section 2. Section 3 briefly describes the architecture of NCDW. In Section 4, we have analyzed the data structures of the underlying operational database and also optimized our dimensional data model for the NCDW storage format. Section 5 shows the experimental results with evaluation and discussion of the model. Finally, Section 6 concludes the paper with a future directions and recommendations.

2. Related work

If we examine the existing national health data warehouse, the World Health Organization (WHO) maintains a centralized health data warehouse for the European Union (WHO/EU, 2012). The U.S. government's Centers for Disease Control and Prevention is holding a National Center of Health Statistics Data Warehouse (NCHS, 1960) based on the following subjects of (1) Government Documents (U.S.); (2) Health and Medicine; and (3) Public Health. As a developing country, the government of Malaysia developed the Malaysian Health Data Warehouse (MyHDW) from 2011 through 2013 (Ministry of Health Malaysia, 2013). In Sæbø, Kossi, Titlestad, Tohouri, and Braa (2011), authors compare different strategies of standardization and methodology of already developed national health data warehouse of four African countries, e.g., South Africa, Zanzibar, Sierra Leone, and Botswana. Among them, South Africa's bottom-up approach is highly optimized, as claimed.

In two fundamentally different data warehouse architecture, Bill Inmon's hub-and-spoke architecture (top-down) (Inmon, 1996) and Ralph Kimball's data mart bus architecture with conformed dimensions (bottom-up) (Kimball & Ross, 2011) are still considered the most popular data warehouse architectures. But which one is the best and most successful among them and others alternative architecture? A web-based survey has been conducted to measure the success value based on the matrices of information quality, system quality, individual impacts, and organizational impacts. Authors found that Kimball's bus architecture has the highest average scores among five data warehouse architectures (Ariyachandra & Watson, 2006). A comparison between data-driven, goal-driven, and user-driven data warehousing methodologies has been evaluated using various assessment criteria. Authors demonstrated that the data-driven and goal-driven approaches are compatible with parallel existence, and by running in parallel, the benefit is higher than using a single methodology (List, Bruckner, Machaczek, & Schiefer, 2002).

Table 1. Related works TAXONOMY of the applied clinical informatics.

Characteristics	**Related Work**
	— Discuss on the capability of CDSS for exploring the potentials of Artificial intelligence through integrating EHR (Shortliffe & Sepúlveda, 2018)
	— CDSS verification, development, and validation (Wasylewicz & Scheepers-Hoeks, 2019)
Decision Support	— Cloud based CDSS for Deep Neural Network classifier (Lakshmanaprabu, Mohanty, Krishnamoorthy, Uthayakumar, Shankar, et al., 2019)

	— Interpretable fuzzy rule bases for data-driven clinical decision support (Chen et al., 2021)
	— Develope HealthMiner, a knowledge discovery tool, and CliniMiner, a predictive analysis and pattern discovery tool based on EHR data (Mullins et al., 2006)
	— Discover similarities and relationships between data elements in large medical records including laboratory data (Harrison Jr, 2008)
Knowledge Discovery	— A comprehensive book on clinical data mining (Epstein, 2009)
	— Data analytics, including data mining, to analyze and utilize the patterns and relationships for clinical decision support (Raghupathi et al., 2010)
	— A review on the foundation principles of mining clinical datasets (Jacob & Ramani, 2012)
	Proposed a new framework of temporal and predictive analysis (Mantovani, 2019)
	— Classification and regression on diabetes data warehouse (Breault, Goodall, & Fos, 2002)
	— Statistical analysis to discover associated disease-finding in a clinical data warehouse (Cao, Markatou, Melton, Chiang, & Hripcsak, 2005)
Disease Analysis	— Identify new risk factors of diseases from EHR (Melamed, Khiabanian, & Rabadan, 2014)
	— UPhenome, a probabilistic graphical model for large-scale discovery of computational models of disease, or phenotypes from heterogenous EHR (Pivovarov et al., 2015)
	— CDSS for disease prediction from complex clinical data (Hashi, Zaman, & Hasan, 2017)
	— Heart disease prediction from EHR (Mohan, Thirumalai, & Srivastava, 2019)
	— A review on PHR, EMR and EHR in the field of public health (Heart, Ben-Assuli, & Shabtai, 2017)
Public Health	— Review the uses of big data on common public health research and practice, while contributing to knowledge, infrastructure, and methodologies and retaining a commitment to the ethical use of data (Mooney & Pejaver, 2018)
	— Leverage EHR Data on public health and population-wide research (McCormick, 2018)

NCDW can be used in many ways to improve national health standards, provide better and prompt services to the patients, and facilitate health-related research among doctors, practitioners, and university researchers. A taxonomy of related work in the applications and usability of the clinical data warehouse is shown in Table 1. There are many ways to derive interesting patterns from clinical data warehouses using various data mining algorithms like association or clustering (Han, Kamber, & Pei, 2011). A new data mining approach defined as HealthMiner has been developed for Clinical Mining, Predictive Analysis, and Pattern Discovery by using clinical records of 667,000 patients' data repository (Mullins et al., 2006). In Prather et al. (1997), authors used 25 years of TMR clinical data of Duke University to create a clinical data warehouse and exploratory factor analysis using data mining techniques to predict and prevent preterm birth and adverse health outcomes. In Wisniewski et al. (2003), the authors provide the importance of a clinical data warehouse to monitor antimicrobial resistance, measure antimicrobial use, detect hospital-acquired bloodstream infections, measure the cost of infections, and detect antimicrobial prescribing errors. The authors provide a review on the benefits of clinical data warehouse applications in creating intelligence for disease management programs, trend analysis, and conducting cohort researches to improve the health of populations (Karami, Rahimi, & Shahmirzadi, 2017).

Table 2. A comparison on the features in existing health informatics systems of Bangladesh with the NCDW.

System	Record Linkage	Privacy	Interoperability	Supporting Data	Decision Support	Disease Management	Population Wide Research	Knowledge Discovery	Research Dataset
National COVID-19 Dashboard	No	Yes	National e-Health standard	COVID-19 test records	COVID-19	COVID-19	Yes	NA	Yes
DHIS2 Central Database	No	Yes	National e-Health standard	Summary data of govt. hospital (irregular data input)	No	No	No	No	No
DHIS2 System (Individual records)	No	Yes	National e-Health standard	Summary data from Community Clinic, Union level facilities and Community (irregular data input)	No	No	No	No	No
National Real-time Health Informatics System - DGHS	No	Yes	NA	Mortality, Fertility, Antenatal, Contraception, Pneumonia, Diarrhea, and Govt. healthcare performance data	Yes	No	Yes	No	No
National Surveillance System	No	Yes	NA	Cervical and Breast Cancer, VPDs and AEFI, COVID-19, Maternal Health & Child Immunization	Yes	No	Yes	No	Yes
Public and Private Healthcare Informatics System	No	Yes	HL7	PHR, EMR, EHR	No	No	No	No	No
NCDW	Yes	Yes (HIPAA)	HL7, SNOMED CT, LOINC, ICD-11	Pathology, Radiology, Genomics, Molecular, Behavior, Lifestyle, Ambient, Summary data	Yes	Yes	Yes	Yes	Yes

The Government of Bangladesh has already implemented numerous health informatics & surveillance systems; a central database for summary data under the Directorate General of Health Service (DGHS) is praise-worthy. However, the lack of record linkage, standardization & interoperability, data quality, analytical ability, etc., is intercepting the process of population-wide research, disease analysis and prevention planning, effective decision making, knowledge discovery, and data-driven research. A comparison of the supported features of different national and other health informatics systems in Bangladesh shows in Table 2. The extensive experimental studies that incorporated our previous research findings have resolved underlying difficulties in developing a national clinical data warehouse.

3. NCDW architecture

We follow the data-driven bottom-up approach to integrate clinical data into NCDW and provide analytical solutions supported by the data. The architecture of the NCDW model based on the Kimball's data mart bus architecture with conformed dimensions is illustrated in Fig. 1. The five stages of the framework briefly describe below.

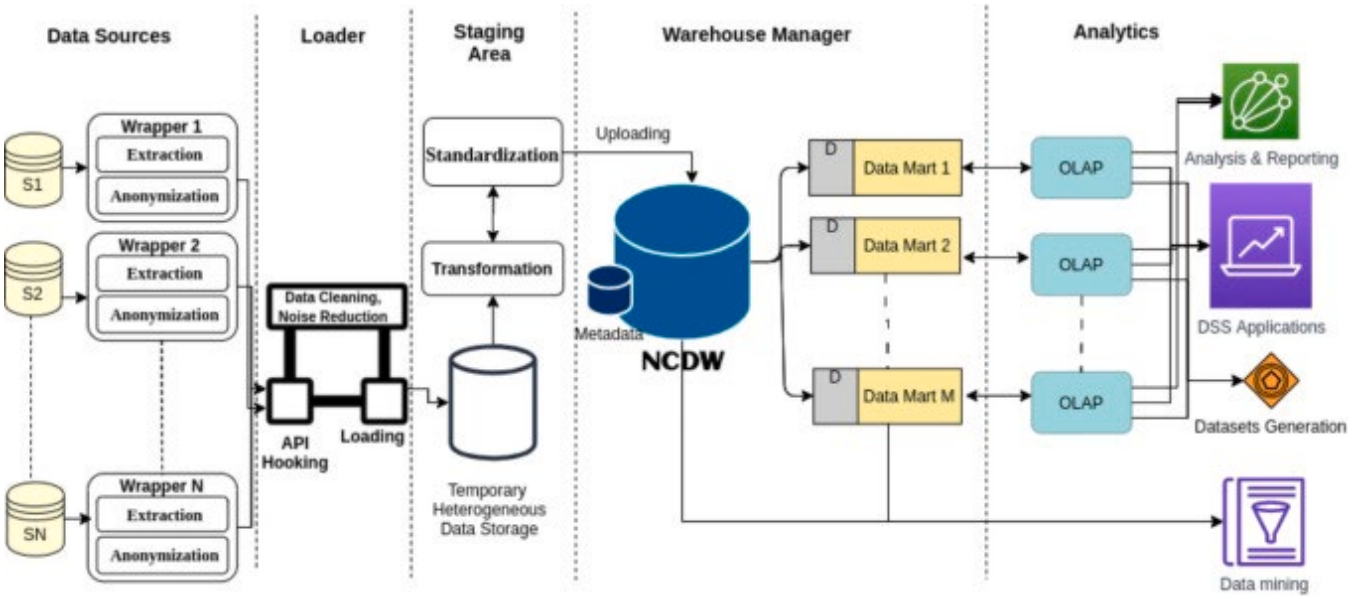

Fig. 1. Wrapper-based architecture of the National Clinical Data Warehouse of Bangladesh.

3.1. Wrapper-based anonymous data acquisition

Clinical data from different government and private sources such as hospitals, clinics, diagnostic centers, research centers are collected in this module. Each data endpoint is the operational databases of the Hospital Information System(HIS), Laboratory Information System(LIS), or Diagnosis Information System(DIS). A data source-specific wrapper service is deployed to the operational server to collect data. Each wrapper contained three core components- (1) Data extraction, (2) Data anonymization, and (3) API endpoint to transfer data in a structured format (e.g., JSON). In addition, we perform some essential processes to extract data from the sources that rely on the source structure. For instance, structured data, e.g., relational or non-relational databases, requires database connectors, data tables, and columns selection and filtering to extract data. Whereas, semi-structured or unstructured data, e.g. XML, text, there will be a need for important source file selection, data parser, feature selection, and data filtering to extract data. Nevertheless, there is no proven evidence that clinical data has been stored in a semi-structured or unstructured format. The customization option in a base wrapper for individual wrapper design minimized the complexity to extract and transfer anonymized data in a unified and uninterrupted manner.

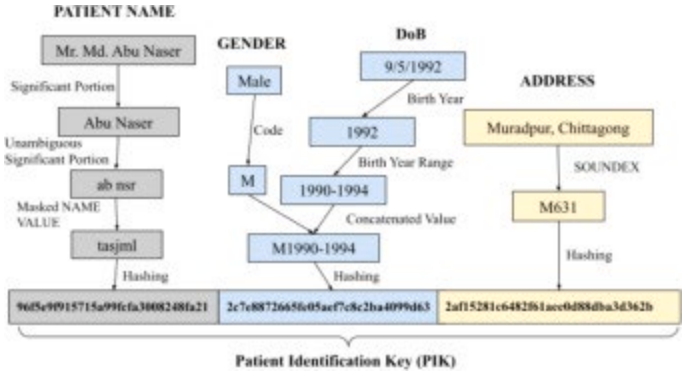

Fig. 2. Illustration of a PIK generation using KSRL algorithm .

3.1.1. Data anonymization, privacy & security

The purpose of data anonymization is to preserve the privacy of a patient's personal information. The addressable and required HIPAA guidelines have been followed to protect the privacy of PHI, patient safety, and system security. Key-based Secured Record Linkage (KSRL) (Khan & Hoque, 2019) algorithm has been used to transform the identifiable patient information into non-identifiable anonymous data and incrementally link patient records into the warehouse. How the KSRL algorithm generates a Patient Identification Key (PIK) from the PHI is illustrated for better understanding in Fig. 2. The algorithm takes *PATIEN NAME, GENDER, DoB, and ADDRESS* as input and return a combined hashed value. The steps of the anonymized record linkage process using the KSRL algorithm are as follow:

–**In step 1,** an encrypted PIK based on KSRL is generated for each patient record using available patient identifiable data.

–**In step 2,** all identifiable data that capable of identifying individual patients are removed from the health record.

Privacy evaluation of the KSRL algorithm using frequency analysis found that the re-identification of the patients from PIK is not possible. The anonymous clinical records of corresponding PIK have been transferred through APIs hooked with the loader of NCDW. A private-key-based encryption protocol has been used to avoid data leaks at the time of data transfer through the network. JSON Web Token (JWT) (Jones, Campbell, & Mortimore, 2015) based authentication is used to access the API endpoint of a wrapper securely.

3.2. Data loading and staging

Heterogeneous data from different sources are temporarily loaded into a Temporary Heterogeneous Data Storage of the staging area. Some essential steps have been performed to control and validate data storage in a unified format, such as:

1. 1.Different data preprocessing techniques have been performed for data cleaning, noise reduction, duplicate, inconsistent, and wrong value handling. Such as, we have applied lower-casing, special character removal, normalization, etc., in textual data. In the case of numerical data, we have applied the Min–Max Normalization technique to normalize the result — for instance, Blood glucose level (mmol/L) normalization.
2. 2.Single Center Imputation from Multiple Chained Equation (SICE) algorithm (Khan & Hoque, 2020b) has been used for the missing value imputation.

A mediator has been designed by satisfying the fundamental functionalities. Such as wrapper API integration using a concrete procedure, data visualization and controlling data flow parameters, automated and manual

data mapping options, metadata management, scheduler, refresh, and load, etc. ensures the consolidation of any source wrapper without interruption.

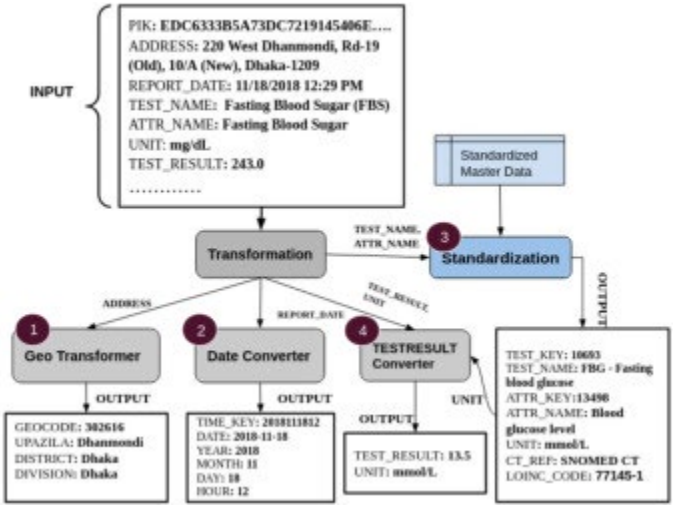

Fig. 3. Illustration of the transformation and standardization process .

3.3. Transformation, standardization, and uploading to the NCDW

In a data-driven system, the strategies and techniques to improve data quality are Acquisition of new data, Source trustworthiness, Record linkage, Standardization, etc (Batini, Cappiello, Francalanci, & Maurino, 2009). Data quality control is highly resting on the data transformation and standardization in a definite format.

A set of transformation rules has been performed to format and convert several data fields, e.g., date, geocode, measurement unit. Transformation rule includes- (1) Conversion and denormalization operate on both storage and measuring units to assemble data in a uniform format; (2) Matching and mapping associates equivalent fields in different sources with the master data, e.g., address mapping with GEOCODE; (3) Converting clinical test, test attribute, reference result, etc., into standardized data.

Data standards are the principal informatics component necessary for information flow through the national health information infrastructure. Common data standards support the broad scope of data collection and reporting requirements, effective assimilation of new knowledge into decision support tools, and support data exchange. As per the NIH patient safety framework (Erickson, Wolcott, Corrigan, Aspden, et al., 2004), standardized healthcare data must satisfy the following criteria.

I. ***Definition of data elements***- Includes standard objects description like Patient Identification Key, gender, and Date of Birth; diagnosis; hospital information; geography; laboratory results and report delivery date. Specific clinical information: blood glucose level or cholesterol level etc. Universal data types: date, time, numeric, currency, or coded elements rely on terminologies. And *UCUM* the standard scientific units.

II. ***Data interchange formats*** - Includes Message Format Standards, Document Architecture, Clinical Templates, User Interface, Patient Data Linkage. As the measurement of laboratory tests in different healthcare of Bangladesh ensues on HL7, the international standard of data exchange format. Hence, we have designed the architecture and functional requirements of the data warehouse by following HL7 EHR Clinical Research Functional Profile (CRFP).

III. ***Terminology*** - Terminology is one of the most important elements for standardizing health information that enables the interoperability and knowledge representation as accepted internationally, which is missing in the existing ehealth standards and interoperability framework of Bangladesh (Directorate

General of Health Services, 2012) for the clinical reference. NCDW utilizes the **SNOMED CT**(*e.g. SCTID, Term.*) and **LOINC** (*e.g. LOINC CODE, Long Common Name, Component, Class, UCUM Units.*) to determine the standard clinical terms, and measurement procedure.

IV. ***Knowledge representation*** - Finally, the knowledge representation considers as the standard method for electronically representing medical literature, clinical guidelines, and the standard for decision support. The elements of external resources used for the NCDW knowledge representation are described below:

- **NHS Pathology**- In Bangladesh, medical practitioners and doctors, as well as medical curriculum, follow UK standards. Clinical reference solely based on NHS guidelines and books. Therefore, we incorporate the NHS pathology handbook for the following purposes: (1) Laboratory wise test classification; (2) Identify the categories of laboratory tests; and (3) Reference ranges (Includes- *Reference type, Age, Gender, Lower limit, Upper limit.)*
- **ICD-11**- We have constructed a tree of used diseases based on the ICD-11 disease classification where each node contains a disease class. The elements of a disease node are *ICD-11 Entity id as IDENTITY, Disease name, Parent disease node, ICD-11 disease code as CODE, Definition of disease, Reference, and URL.*
- **LAB Tests Online and relevant resources**- To be used for the purpose of associated laboratory test mapping with ICD-11 disease classes.

The process of transformations and standardization is illustrated in Fig. 3 with a paradigm. Input data contains staging data of an anonymous record. The steps of transforming that data are as follow:

1. **Geo Transformer:** The hierarchy of the geography data is *Division→ City→ District→ Upazila*, and there are a total of 556 GEOCODE in the lowest level. If the INPUT geographic data contains the hierarchy, then the lowest level is mapped with the GEOCODE. Otherwise, the ADDRESS of the INPUT data is mapped with the GEOCODE based on the cosine similarity score of Deep Structured Semantic Models (DSSM) (Hu, Dang, & Tan, 2019). For instance, the split ADDRESS *['220', 'West', 'Dhanmondi', 'Rd-19(Old)', '10/A(New)', 'Dhaka-1209']* mapped with "Dhanmondi" from the hierarchy of geography based on highest score. In case of the absence of geographic data, healthcare location is considered the patient's geography.
2. **Date Converter:** The INPUT date might be any format like yyyy-mm-dd hh:mm, dd/mm/yyyy hh:mm:ss.s, or dd-mmm-yyyy. Date converter converts all possible date format into a specific format (e.g., DOB format — dd-mm-yyyy, reporting date format — dd-mm-yyyy hh).
3. **Standardization:** Standardized master data of 1135 diagnostic tests with a coding system have been constructed and cross-validated by domain experts. Here we have utilized the internationally accepted standard of interoperability described above. Diagnosis tests of a source are mapped one-time with the standardized laboratory tests. This map is used for maintaining the consistent clinical data integration as per the standard reference.
4. **TESTRESULT Converter:** The output *Unit* of the standardization step has been used to convert the result of a laboratory test into a unified measurement unit. For instance, the Blood Glucose Level of 243.0 mg/dL converted to 13.5 mmol/L based on the output of the standardization step.

At last, the transformed and standardized data of the staging area merged and uploaded to the central storage of NCDW using batch processing.

3.4. Warehouse management and monitoring

A multidimensional data model (describe in Section 4) has been used to store data that is efficient for complex query operations, e.g., cube, dicing, slicing, pivoting, roll-up, drill-down. The NCDW management

module performs- (1) the controlling process like transformation, standardization, merging the source data of the temporary store into the NCDW; (2) operational, technical, process execution, and other metadata management; (3) master data management includes disease classification tree, clinical test, attributes, and reference result, geographical hierarchy; (4) automating data mart fact generation and controlling data mart pipeline; (5) analyzing query profiles; (6) creating indexes, business views, partition views against the base data; (7) data backup and recovery options. Furthermore, monitoring data loads, response time, queries and reports, data archives, backups/recoveries, statistical inference of data usages, users, and activities are also managed in this module.

3.5. Analytics and applications

NCDW has the vision to develop a next-generation clinical analytic and research data platform, will enable us to achieve a broad range of applications by utilizing the knowledge and insight of the data. We have developed an analytical dashboard using the Python flask backend framework, and Angular 11 client-side programming (eUbiMy-Soft, 2021). Authorized end users can browse the web-based semantic analytical dashboard of NCDW for multi-purposes.

3.5.1. Online analytical processing (OLAP)

We used Apache Superset (www.superset.apache.org), an open-source cloud-native application for data exploration and data visualization. An extensive OLAP server for each data mart has been configured in the superset to understand the data insight and business intelligence, analysis, and reporting purposes.

3.5.2. Data mining and applications

NCDW will store enormous data estimated at 6.23GB/day if we integrate minimum possible 8717 data sources (Khan & Hoque, 2015). This vast data can prompt many interesting patterns to improve national health services using different data mining techniques. A suitable read-only exploratory interface will be provided by connecting with the central data warehouse or any specific data mart to apply data mining, machine learning, or other algorithms.

3.5.3. Dataset generation

Researchers need health datasets of specific diseases or diagnoses. Anonymous research dataset will be generated based on different perspectives and provided through API or any machine-readable format file (e.g., CSV, Excel) to the authorized/requested user.

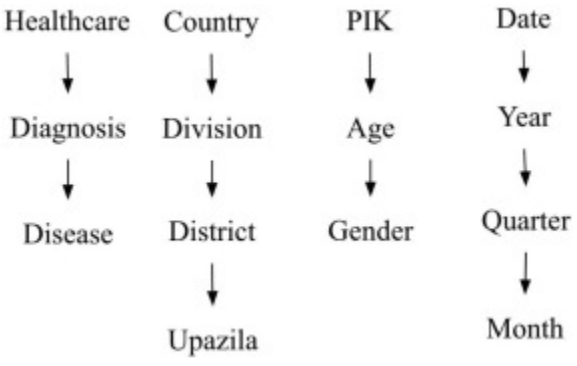

Fig. 4. CUBE hierarchy in disease analysis .

4. Data modeling and analysis

4.1. Operational data analysis

An investigation in the underlying operational database has been performed and checked whether the data required to create the clinical data warehouse is available. We have found that the structure of database tables and entity relationships are primarily identical for the three investigated information systems. The selected operational data are grouped into three categories.

(1) *Healthcare information.* Includes healthcare profile, location, branch, contacts, laboratories;
(2) *Patient details.* Registered outpatient and inpatient information including name, gender, date of birth, contact, address, registration date, etc., and invoice details;
(3) *Laboratory test records.* Includes diagnosis test items containing test name, specimen, lab, price, etc., test group, measure attributes of the test consists of attribute name, unit, default value, analyzer, etc. and the measure attribute result against the patient invoice. As an example, the measure attributes of *Lipid Profile* diagnostic test are *Total Cholesterol, HDL, LDL,* and *Triglycerides*;

The statistical analysis of laboratory information from the sources indicated that an average of 66.9% of the tests was utilized. Among them, 77.1% are identical and interchangeable. Therefore, the quality of the data is highly dependent on the associated clinical test matching for consistency, where the data will come from diverse sources.

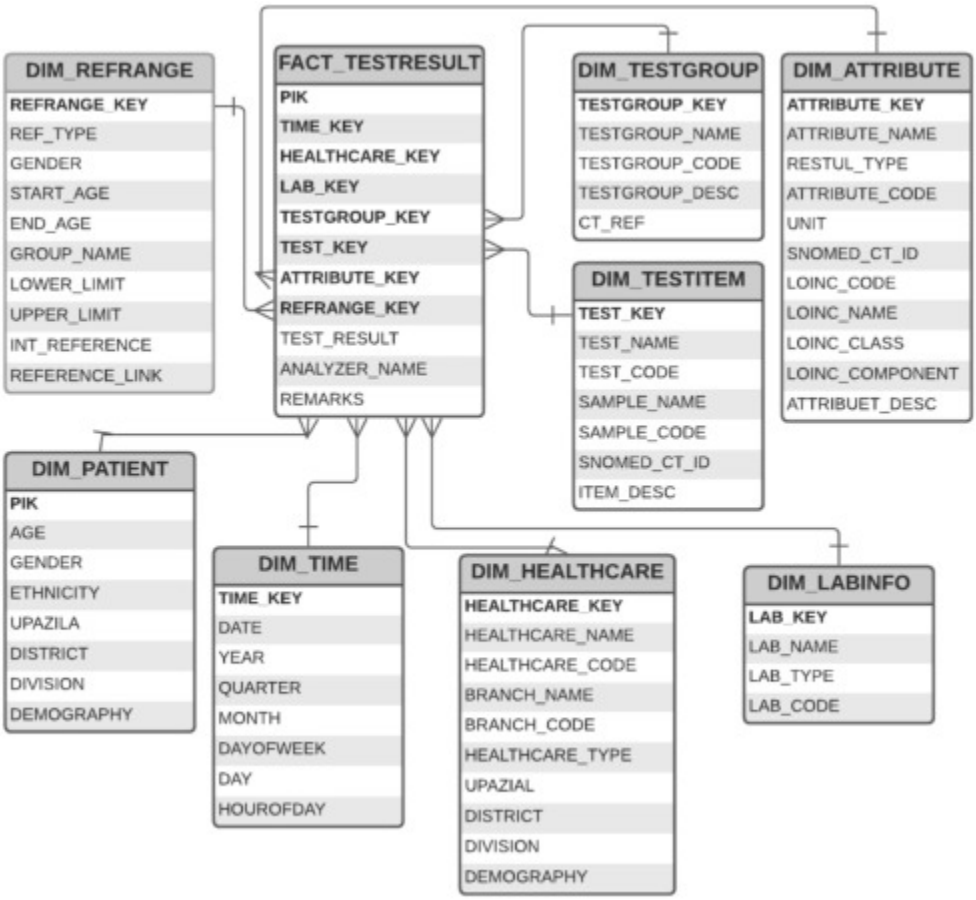

Fig. 5. Central schema of NCDW for laboratory records storage.

4.2. Analysis of dimensional data modeling

The most popular data model for a data warehouse is the multidimensional data modeling, which can exist in the form of a star schema, a snowflake schema, or a fact constellation schema (Linstedt & Olschimke, 2015). A star schema of NCDW with eight dimension tables based on the analysis of the reconciled data is shown in Fig. 5. The fact table *TESTRESULT* is considered along all eight dimensions: *PATIENT*, *TIME*, *HEALTHCARE*, *LABORATORY*, *TESTGROUP*, *TESTITEM*, *ATTRIBUTE*, and *REFRANGE* and contains keys to each of the eight dimensions, along with three measures: test_result as nonadditive fact, and analyzer_name, and remarks as factless fact. The result of a clinical laboratory test contains the value of measure attributes that are used to diagnose and screen diseases. Measurement value and unit might differ based on the hardware analyzer used in healthcare and remarks containing important interpretations, notes, and comments. Hence, we have stored the factless attribute in the central fact table. To minimize the size of the fact table and computational complexity, dimension identifiers such as healthcare_key, lab_key, testgroup_key, testitem_key, attribute_key, and refrange_key are system-generated five-digit identifiers. The time_key is the integer format of DateTime (e.g., time_key of 19941108 3:55 PM is 1996110815), and the PIK generated through the data anonymization process in the wrapper. The *PATIENT* dimension is only stored the non-identifiable patient data to protect the patient safety, as described in Section 3.1.1.

Some constraints (e.g., geographical hierarchy) may introduce redundancy in the order where the snowflake schema reduces redundancy. Still, it is not as popular as the star schema in data warehouse design (Chen, 2003). The aggregated data generated from this model is appropriate for CUBE creation, as shown in Fig. 4, which can be used for diagnosis and disease analysis. Furthermore, the conformed dimensions enfranchise the schema to extend the NCDW to merge genomic/molecular data through to behavioral/lifestyle data by adding new fact and dimension tables in the same fashion.

Table 3. Details of the derived data marts.

Basic Info	**Laboratory Tests**	**Measures**	**Dimensions**
Name: DENGUE Disease class: Dengue ICD-11 Entity ID: 1959883044	Nucleic acid, amplification tests (NAATs), Dengue Antibody IgG & IgM	*Test Result* *Dengue status*	*PATIENT, TIME, HEALTHCARE, LAB, TESTITEM, ATTRIBUTE, REFRANGE*
Name: DIABETES Disease class: Diabetes mellitus ICD-11 Entity ID: 465177735	Random blood sugar, Urine glucose test, Fasting blood sugar, HbA1c test, Oral glucose tolerance tests, C-peptide tests, Lipid profile	*Test Result* *Normal status* *Prediabetes status* *Diabetes status*	*Same as avobe*

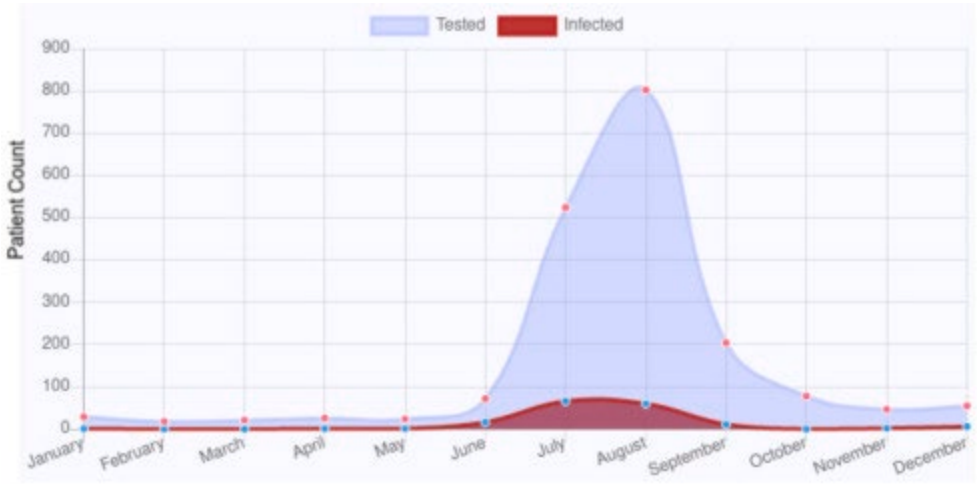

Fig. 6. Dengue outbreak season detection.

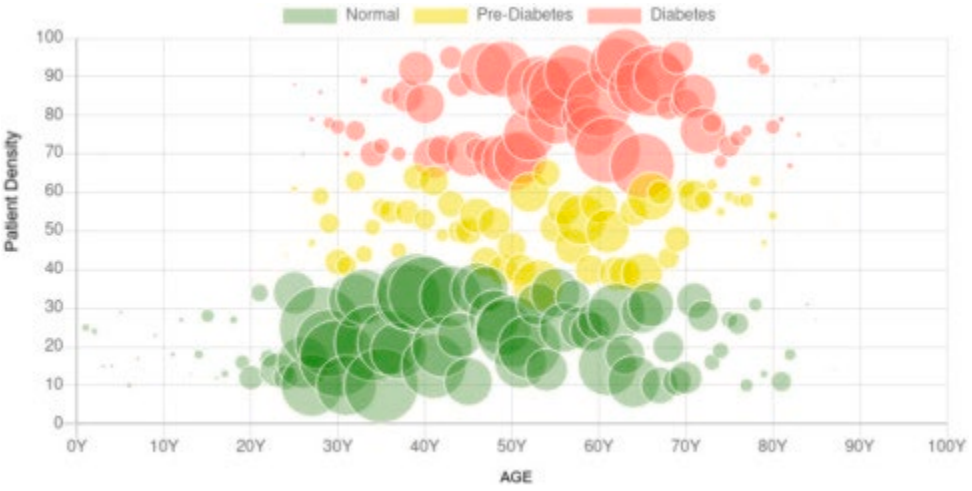

Fig. 7. Diabetes at-risk age group detection.

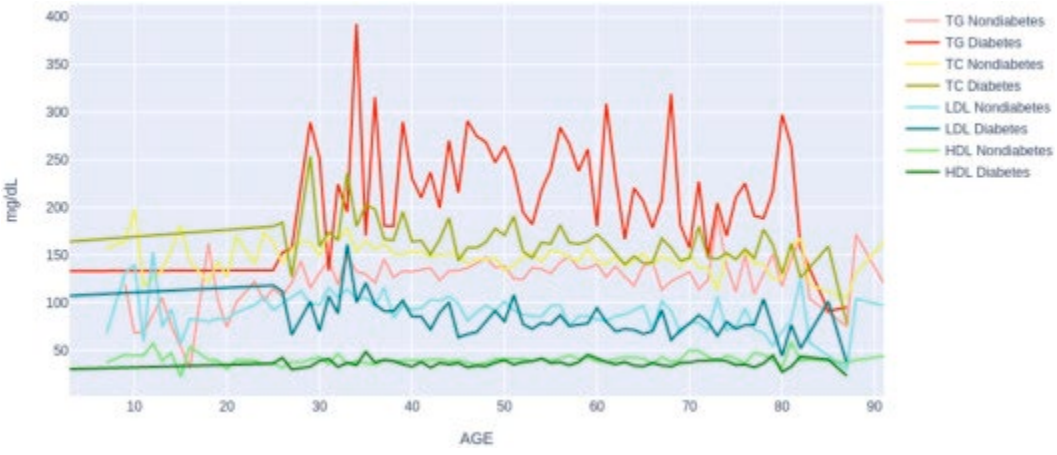

Fig. 8. Average lipid level for Diabetes and Non-diabetes patient with respect to AGE.

5. Experimental result and discussion

We have developed a prototype of NCDW and deployed it in a server of IICT, BUET; server configurations are CPU: 2*4, HDD: 200 GB, and RAM: 32 GB. Experimentally, 1 161 654 clinical records of 9443 anonymous patients

from three hospitals have been loaded into the PostgreSQL database by completing all required processes and steps.

By considering the high velocity of clinical data, Apache Hadoop, a distributed file system, has been configured as a single node cluster and stored the Heterogeneous Temporary Data, which can be scaled for a large population by adding more DataNodes. 54% storage space is reduced when the staging data has been uploaded to the central star schema.

Two subject-oriented data marts (DENGUE and DIABETES) have been derived from the NCDW, detailed is shown in Table 3. The *status* measures in the data marts have been calculated by encoding categorical data or using the reference range as a cut-off value. Some complex CUBE, slice, dice, drill-down, roll-up queries have been performed in the data marts and visualized the analytical outcomes in enhanced graphs. In Fig. 6, the outbreak season of dengue infection has been detected and found that **June–September** months are the most critical time for dengue outbreak in the demography of Bangladesh. An experimental analysis of 40,476 dengue cases of Bangladesh occurring during 2000–2017 showed that 49.73% dengue infection occurred during the monsoon season (May–August) and 49.22% during the post-monsoon season (September–December) (Mutsuddy, Tahmina Jhora, Shamsuzzaman, Kaisar, & Khan, 2019), which aligned with our findings.

Similarly, the outcomes from the DIABETES data mart have been evaluated in Table 4. There are hundreds of disease classes in ICD-11 that can be derived from NCDW as a data mart, such as (a) Certain infectious or parasitic diseases (e.g., COVID-19, malaria, HIV, chikungunya, viral hepatitis, influenza). (b) Diseases of the blood or blood-forming organs (e.g., anemias, thalassemia, pure red cell aplasia, thrombophilia). (c) Endocrine, nutritional or metabolic diseases (e.g., diabetes mellitus, insulin resistance, lipidoses). (d) Diseases of the circulatory system. (e) Diseases of the respiratory system, etc.

Table 4. Evaluation of the result of Figs 7 and 8.

NCDW Result	**Comparison-1**	**Comparison-2**
Diabetes developing age is 27 or older and more likely developed at age 40 or older (Fig. 7)	Type 2 diabetes develop at age 45 or older according to [59]	In [72], 10.3% of men and 9.6% of women in Europe aged 25 years and over are diabetes patient
Diabetes patient's avg. TG >150mg/dL and adult age 25–40 has peaked TG (Fig. 8)	Non-diabetes patient's TG levels are <150 mg/dl in [4]	Trends in adults aged 20 and over with elevated TG in [10]

	Age	Gender	Date	Healthcare	Test	Attribute	Result	Unit
0	60.0	F	2020-02-22	100001	FBG - Fasting blood glucose	Blood glucose level	6.0	mmol/L
1	60.0	F	2020-02-22	100001	HbA1c	HbA1c	6.7	%
2	60.0	F	2019-11-03	100002	FBG - Fasting blood glucose	Blood glucose level	6.0	mmol/L
3	60.0	F	2019-11-03	100002	HbA1c	HbA1c	6.5	%
4	59.0	F	2019-08-01	100001	HbA1c	HbA1c	6.2	%
5	59.0	F	2019-08-01	100001	FBG - Fasting blood glucose	Blood glucose level	5.2	mmol/L
6	59.0	F	2019-06-27	100002	HbA1c	HbA1c	6.2	%
7	59.0	F	2019-06-27	100002	FBG - Fasting blood glucose	Blood glucose level	5.0	mmol/L
8	59.0	F	2019-05-06	100001	FBG - Fasting blood glucose	Blood glucose level	5.7	mmol/L
9	59.0	F	2019-04-01	100001	FBG - Fasting blood glucose	Blood glucose level	6.1	mmol/L
10	59.0	F	2019-03-27	100001	RBG - Random blood glucose	Blood glucose level	7.2	mmol/L
11	59.0	F	2019-03-01	100001	FBG - Fasting blood glucose	Blood glucose level	5.5	mmol/L
12	59.0	F	2019-01-13	100002	HbA1c	HbA1c	5.9	%
13	59.0	F	2019-01-13	100002	FBG - Fasting blood glucose	Blood glucose level	5.8	mmol/L
14	59.0	F	2018-12-14	100001	FBG - Fasting blood glucose	Blood glucose level	6.0	mmol/L

Fig. 9. Diabetes diagnosis tests records of a patient (PIK is EDC6333B5A73DC7219145406E4BEE....3D362B) retrieve from DIABETES data mart.

The linkage records from numerous sources (see in Fig. 9) can improve the performance of diagnostic and prognostic machine learning algorithms, fueling observational research and improving clinical decisions at the point-of-care (Seneviratne, Kahn, & Hernandez-Boussard, 2018). Furthermore, In the progression of population-wide analysis and study, the data of NCDW is capable of utilizing the developed techniques and algorithms in the field of applied clinical informatics, such as Insight into disease processes, Intelligence in disease management, Understanding of current clinical practice, Public health issues, Evidence-based practice guidelines, Protocol development, Care standards development, Outcomes measures, Wellness management, Clinical risk adjustment, etc (Evans et al., 2012, Krasowski et al., 2015, Shahid et al., 2021).

6. Future works and conclusion

A novel NCDW framework of Bangladesh by overcoming the challenges to develop a clinical big data platform has been presented in this paper. The proposed wrapper-based architecture is effective for the developing countries to integrate and standardize disperse clinical data with minimal cost and shortest time. HIPAA compliance, interoperability, consistency in data integration, data quality control, and uniformity are the most potent components of the framework. The analytical efficacy from the derived data marts found to be worthwhile, as there is no existing centralized data repository in Bangladesh to support the features of NCDW.

NCDW will be enriched by integrating radiology, genomic, molecular, socio-economic factors, behavior & lifestyle factors, and ambient data in the future. The diverse characteristics and consistent linkage records can optimize descriptive, diagnostic, predictive, and prescriptive analysis of disease research. In Bangladesh, the medical practitioner uses the UK standard for clinical reference that might be different for this demography, which can be figured out by data-driven research. Furthermore, the linkage records can be used for developing a personalized patient care system with the patient's consent to support physician decisions, disease-risk profiles, and recommendations. Besides, a disease-centered knowledge base development from the NCDW can empower predicting health-risk and corresponding wellness plans. In addition, we can construct homogeneous or

heterogeneous information networks for high disease-risk community detection or searching in public health analysis.

In Bangladesh's existing healthcare systems, many vital records such as physiological conditions, behavioral, social, and lifestyle factors are absent as electronic records. Therefore, we recommend storing essential behavioral and lifestyle data (e.g., height, weight, blood pressure, physical activity, sleeping disorder, tobacco smoking) in the exiting operational system significantly impacts clinical research.

Acronyms

CDSS Clinical Decision Support System
HER Electronic Health Records
EMR Electronic Medical Records
PHR Personal Health Records
PHI Protected Health Information
NCDW National Clinical Data Warehouse
HIPAA Health Insurance Portability and Accountability Act
BMRC Bangladesh Medical Research Council
HL7 Health Level 7
SNOMED CT Systemized Nomenclature of Human and Veterinary Medicine, Clinical Terms
LOINC Logical Observation Identifiers, Names, and Codes
ICD-11 International Classification of Diseases - 11
DGHS Directorate General of Health Services
TC Total Cholesterol
LDL Low-Density Lipoprotein
HDL High-Density Lipoprotein
TG Triglycerides
NHS National Health Service
IICT Institution of Information and Communication Technology
eSRD-Lab eSystems Research and Developement Lab

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This study was done by an academia-industry research collaboration among three academic entities: (1) eSRD-Lab, Department of CSE, BUET; (2) Ubicomp Lab, Department of Computer Science, Marquette University; (3) Institute of Information and Communication Technology (IICT), BUET; and two software companies: (1) MySoft Ltd., Bangladesh; and (2) Ubitrix Inc., WI, USA. We would like to thank the Bangladesh Medical Research Council (BMRC) for processing ethical clearance.

References

Al-Zaman, M. S. (2020). Healthcare crisis in Bangladesh during the COVID-19 pandemic. The American Journal of Tropical Medicine and Hygiene, 103(4), 1357–1359.

Alhazmi, F. (2019). The Ethical Challenge of Conflicts of Interest in Healthcare. (Ph.D. thesis), Duquesne University.

Ariyachandra, T., & Watson, H. J. (2006). Which data warehouse architecture is most successful? Business Intelligence Journal, 11(1), 4.

Association, A. D., et al. (2004). Dyslipidemia management in adults with diabetes. Diabetes Care, 27(suppl 1), s68–s71.

Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for data quality assessment and improvement. ACM Computing Surveys, 41(3), 1–52.

Birkhead, G. S., Klompas, M., & Shah, N. R. (2015). Uses of electronic health records for public health surveillance to advance public health. Annual Review of Public Health, 36, 345–359.

Breault, J. L., Goodall, C. R., & Fos, P. J. (2002). Data mining a diabetic data warehouse. Artificial Intelligence in Medicine, 26(1–2), 37–54.

Calman, N., Hauser, D., Lurio, J., Wu, W. Y., & Pichardo, M. (2012). Strengthening public health and primary care collaboration through electronic health records. American Journal of Public Health, 102(11), e13–e18.

Cao, H., Markatou, M., Melton, G. B., Chiang, M. F., & Hripcsak, G. (2005). Mining a clinical data warehouse to discover disease-finding associations using co-occurrence statistics. 2005, In AMIA annual symposium proceedings (p. 106). American Medical Informatics Association.

Carroll, M. D., Kit, B. K., & Lacher, D. A. (2015). Trends in elevated triglyceride in adults: United States, 2001-2012. US Department of Health and Human Services, Centers for Disease Control and .. ..

Chen, Z. (2003). Data warehousing and data marts. In H. Bidgoli (Ed.), Encyclopedia of information systems (pp. 521–533). New York: Elsevier, http://dx.doi.org/10.1016/B0-12-227240-4/00036-8, URL https://www.sciencedirect.com/science/article/pii/B0122272404000368. Smart Health 23 (2022) 10023813 M.R. Mia et al.

Chen, T., Shang, C., Su, P., Keravnou-Papailiou, E., Zhao, Y., Antoniou, G., et al. (2021). A decision tree-initialised neuro-fuzzy approach for clinical decision support. Artificial Intelligence in Medicine, 111, Article 101986.

Daniels, M., & Donilon, T. E. (2017). The emerging global health crisis. Noncommunicable Diseases in Low-and Middle-Income Countries [Homepage on the Internet]. Council on Foreign Relations, 72.

Devi, G., & Rizvi, S. (2022). Artificial intelligence for personalized medicine with EHR and genomic information. In Smart systems: Innovations in computing (pp. 573–582). Springer.

Directorate General of Health Services (2012). Bangladesh ehealth standards and interoperability framework]. Retrieved from URL https://dghs.gov.bd/images/docs/ehealth/standards_{a}nd_{i}nteroperability_{d}ocument_{f}inal_{5}.01.14.pdfAccessed 1 May 2021.

Elliott, T. E., O'Connor, P. J., Asche, S. E., Saman, D. M., Dehmer, S. P., Ekstrom, H. L., et al. (2021). Design and rationale of an intervention to improve cancer prevention using clinical decision support and shared decision making: A clinic-randomized trial. Contemporary Clinical Trials, 102, Article 106271.

Epstein, I. (2009). Clinical data-mining: Integrating practice and research. Oxford University Press.

Erickson, S. M., Wolcott, J., Corrigan, J. M., Aspden, P., et al. (2004). Patient safety: Achieving a new standard for care. National Academies Press, http://dx.doi.org/10.17226/10863.

eUbiMy-Soft (2021). National clinical data warehouse. Retrieved from URL http://103.94.135.217. Accessed 1 May 2021.

Evans, R. S. (2016). Electronic health records: Then, now, and in the future. Yearbook of Medical Informatics, 25(S 01), S48–S61.

Evans, R. S., Lloyd, J. F., & Pierce, L. A. (2012). Clinical use of an enterprise data warehouse. 2012, In AMIA annual symposium proceedings (p. 189). American Medical Informatics Association.

Hagar, Y., Albers, D., Pivovarov, R., Chase, H., Dukic, V., & Elhadad, N. (2014). Survival analysis with electronic health record data: Experiments with chronic kidney disease. Statistical Analysis and Data Mining The ASA Data Science Journal, 7(5), 385–403.

Han, J., Kamber, M., & Pei, J. (2011). The Morgan Kaufmann series in data management systems: vol. 5, Data mining concepts and techniques (3rd ed.). (4), (pp. 83–124).

Harrison Jr, J. H. (2008). Introduction to the mining of clinical data. Clinics in Laboratory Medicine, 28(1), 1–7.

Hashi, E. K., Zaman, M. S. U., & Hasan, M. R. (2017). An expert clinical decision support system to predict disease using classification techniques. In 2017 international conference on electrical, computer and communication engineering (pp. 396–400). IEEE.

Heart, T., Ben-Assuli, O., & Shabtai, I. (2017). A review of PHR, EMR and EHR integration: A more personalized healthcare and public health policy. Health Policy and Technology, 6(1), 20–25.

Hu, W., Dang, A., & Tan, Y. (2019). A survey of state-of-the-art short text matching algorithms. In International conference on data mining and big data (pp. 211–219). Springer.

Humphrey, B. A. (2021). Data Privacy vs. Innovation: A Quantitative Analysis of Artificial Intelligence in Healthcare and Its Impact on HIPAA regarding the Privacy and Security of Protected Health Information. (Ph.D. thesis), Robert Morris University.

Inmon, W. H. (1996). The data warehouse and data mining. Communications of the ACM, 39(11), 49–51.

Jacob, S. G., & Ramani, R. G. (2012). Data mining in clinical data sets: A review. Training, 4(6).

Joarder, T., Chaudhury, T. Z., & Mannan, I. (2019). Universal health coverage in Bangladesh: Activities, challenges, and suggestions. Advances in Public Health, 2019.

Jones, M., Campbell, B., & Mortimore, C. (2015). JSON Web Token (JWT) profile for OAuth 2.0 client authentication and authorization Grants. May-2015.{Online}. Available: https://Tools.Ietf.Org/Html/Rfc7523.

Karami, M., Rahimi, A., & Shahmirzadi, A. H. (2017). Clinical data warehouse: An effective tool to create intelligence in disease management. The Health Care Manager, 36(4), 380–384.

Khan, S. I., & Hoque, A. S. M. L. (2015). Towards development of health data warehouse: Bangladesh perspective. In 2015 international conference on electrical engineering and information communication technology (pp. 1–6). IEEE.

Khan, S. I., & Hoque, A. S. M. L. (2016a). An analysis of the problems for health data integration in Bangladesh. In 2016 international conference on innovations in science, engineering and technology (pp. 1–4). IEEE.

Khan, S. I., & Hoque, A. S. M. L. (2016b). Towards development of national health data warehouse for knowledge discovery. In Intelligent systems technologies and applications (pp. 413–421). Springer.

Khan, S. I., & Hoque, A. S. M. L. (2017). Health data integration with secured record linkage: A practical solution for Bangladesh and other developing countries. In 2017 international conference on networking, systems and security (pp. 156–161). IEEE.

Khan, S. I., & Hoque, A. S. L. (2019). Secured technique for healthcare record linkage. In Proceedings of the 6th international conference on networking, systems and security (pp. 30–36).

Khan, S. I., & Hoque, A. S. M. L. (2020a). Efficient techniques for privacy preserved incremental record linkage of noisy health data. (Ph.D. thesis), Department of Computer Science and Engineering (CSE), BUET.

Khan, S. I., & Hoque, A. S. M. L. (2020b). Sice: An improved missing data imputation technique. Journal of Big Data, 7(1), 1–21.

Khan, S. I., Hoque, A., & Ullah, M. (2016). National health data warehouse Bangladesh for remote health monitoring: Features, problems and privacy issues. In Remote health monitoring workshop: vol. 6.

Khan, S. I., Hoque, A. S. M. L., et al. (2015). Development of national health data warehouse for data mining. Database Systems Journal, 6(1), 3–13.

Khan, S., Hoque, A., et al. (2016c). Digital health data: A comprehensive review of privacy and security risks and some recommendations. Computer Science Journal of Moldova, 71(2), 273–292.

Kimball, R., & Ross, M. (2011). The data warehouse toolkit: The complete guide to dimensional modeling. John Wiley & Sons.

Krasowski, M. D., Schriever, A., Mathur, G., Blau, J. L., Stauffer, S. L., & Ford, B. A. (2015). Use of a data warehouse at an academic medical center for clinical pathology quality improvement, education, and research. Journal of Pathology Informatics, 6.

Lakshmanaprabu, S., Mohanty, S. N., Krishnamoorthy, S., Uthayakumar, J., Shankar, K., et al. (2019). Online clinical decision support system using optimal deep neural networks. Applied Soft Computing, 81, Article 105487.

Linstedt, D., & Olschimke, M. (2015). Building a scalable data warehouse with data vault 2.0. Morgan Kaufmann.

List, B., Bruckner, R. M., Machaczek, K., & Schiefer, J. (2002). A comparison of data warehouse development methodologies case study of the process warehouse. In International conference on database and expert systems applications (pp. 203–215). Springer.

Mantovani, M. (2019). Approximate data mining techniques on clinical data. (Ph.D. thesis), University of Verona.

McCormick, E. V. (2018). Public health and population health: Leveraging electronic health record data for local population health surveillance. Anschutz Medical Campus: University of Colorado Denver.

Melamed, R. D., Khiabanian, H., & Rabadan, R. (2014). Data-driven discovery of seasonally linked diseases from an electronic health records system. BMC Bioinformatics, 15(6), 1–10.

Ministry of Health Malaysia (2013). Malaysian health data warehouse (MyHDW); publication of ministry of health, Malaysia. Retrieved from URL https: //www.moh.gov.my/moh/images/gallery/publications/myhdw%202011-2013.pdf. Accessed 1 May 2021.

Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. IEEE Access, 7, 81542–81554.

Mooney, S. J., & Pejaver, V. (2018). Big data in public health: Terminology, machine learning, and privacy. Annual Review of Public Health, 39, 95–112.

Mullins, I. M., Siadaty, M. S., Lyman, J., Scully, K., Garrett, C. T., Miller, W. G., et al. (2006). Data mining and clinical data repositories: Insights from a 667,000 patient data set. Computers in Biology and Medicine, 36(12), 1351–1377. Smart Health 23 (2022) 10023814 M.R. Mia et al.

Musen, M. A., Middleton, B., & Greenes, R. A. (2021). Clinical decision-support systems. In Biomedical informatics (pp. 795–840). Springer.

Mutsuddy, P., Tahmina Jhora, S., Shamsuzzaman, A. K. M., Kaisar, S., & Khan, M. N. A. (2019). Dengue situation in Bangladesh: An epidemiological shift in terms of morbidity and mortality. Canadian Journal of Infectious Diseases and Medical Microbiology, 2019.

NCHS (1960). National center for health statistics data warehouse (NCHS). Retrieved from URL http://www.cdc.gov/nchs/index.htm. Accessed 1 May 2021.

NIH (2015). National institute of diabetes and digestive and kidney diseases. Retrieved from URL https://www.niddk.nih.gov/health-information/diabetes/overview/what-is-diabetes/type-2-diabetes. Accessed 1 May 2021.

Perlman, S. E. (2021). Use and visualization of electronic health record data to advance public health. American Public Health Association.

Pivovarov, R., Perotte, A. J., Grave, E., Angiolillo, J., Wiggins, C. H., & Elhadad, N. (2015). Learning probabilistic phenotypes from heterogeneous EHR data. Journal of Biomedical Informatics, 58, 156–165.

Prather, J. C., Lobach, D. F., Goodwin, L. K., Hales, J. W., Hage, M. L., & Hammond, W. E. (1997). Medical data mining: Knowledge discovery in a clinical data warehouse.. In Proceedings of the AMIA annual fall symposium (p. 101). American Medical Informatics Association.

Raghupathi, W., et al. (2010). Data mining in health care. Healthcare Informatics Improving Efficiency and Productivity, 211, 223.

Sæbø, J. I., Kossi, E. K., Titlestad, O. H., Tohouri, R. R., & Braa, J. (2011). Comparing strategies to integrate health information systems following a data warehouse approach in four countries. Information Technology for Development, 17(1), 42–60.

Seneviratne, M. G., Kahn, M. G., & Hernandez-Boussard, T. (2018). Merging heterogeneous clinical data to enable knowledge discovery. In BIOCOMPUTING 2019: Proceedings of the pacific symposium (pp. 439–443). World Scientific.

Shahid, A., Nguyen, T.-A. N., Kechadi, M., et al. (2021). Big data warehouse for healthcare-sensitive data applications. Sensors, 21(7), 2353.

Sherimon, V., Puliprathu Cherian, S., Mathew, R., Kumar, S. M., Nair, R. V., Shaikh, K., et al. (2021). Clinical decision support for primary health centers to combat COVID-19 pandemic. In Intelligent systems (pp. 481–490). Springer.

Shortliffe, E. H., & Sepúlveda, M. J. (2018). Clinical decision support in the era of artificial intelligence. Jama, 320(21), 2199–2200.

Wang, D., Wang, L., Zhang, Z., Wang, D., Zhu, H., Gao, Y., et al. (2021). ‘‘Brilliant AI Doctor’’ in Rural Clinics: Challenges in AI-Powered Clinical Decision Support System Deployment. In Proceedings of the 2021 CHI conference on human factors in computing systems (pp. 1–18).

Wasylewicz, A., & Scheepers-Hoeks, A. (2019). Clinical decision support systems. Fundamentals of Clinical Data Science, 153–169.

Webber, T., Santana, J. M., Vermeulen, A. F., & Bowles, J. K. (2020). Designing a patient-centric system for secure exchanges of medical data. In International conference on computational science and its applications (pp. 598–614). Springer.

WHO (2010). Diabetes data and statistics. Retrieved from URL https://www.euro.who.int/en/health-topics/noncommunicable-diseases/diabetes/data-andstatistics. Accessed 1 May 2021.

WHO (2021). Global health workforce alliance: Bangladesh. Retrieved from URL https://www.who.int/workforcealliance/countries/bgd/en/. Accessed 1 May 2021.

WHO/EU (2012). WHO/Europe data warehouse v3 [Internet]. Retrieved from URL https://dw.euro.who.int. Accessed 1 May 2021.

Wisniewski, M. F., Kieszkowski, P., Zagorski, B. M., Trick, W. E., Sommers, M., & Weinstein, R. A. (2003). Development of a clinical data warehouse for hospital infection control. Journal of the American Medical Informatics Association, 10(5), 454–462.

Yang, S., Xiong, H., Xu, K., Wang, L., Bian, J., & Sun, Z. (2021). Improving covariance-regularized discriminant analysis for EHR-based predictive analytics of diseases. Applied Intelligence: The International Journal of Artificial Intelligence, Neural Networks, and Complex Problem-Solving Technologies, 51(1), 377–395.