IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received July 9, 2019, accepted July 18, 2019, date of publication July 24, 2019, date of current version August 9, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2930730

A Framework for Public Health Monitoring, Analytics and Research

FATIMA KHALIQUE[1], SHOAB A. KHAN[1], AND IRUM NOSHEEN[2]

[1]Department of Computer and Software Engineering, College of Electrical and Mechanical Engineering (CEME), National University of Sciences and Technology (NUST), Islamabad 44000, Pakistan
[2]Department of Electrical Engineering, International Islamic University, Islamabad 44000, Pakistan

Corresponding author: Fatima Khalique (fathema.khalique@gmail.com)

ABSTRACT This paper presents a framework for public healthcare data acquisition and management model based on standard protocol for its easy adoption by any country or international health organizations. The model assumes basic digitization of electronic health record (EHR) at basic health facilities. Thus far, the models in the literature have utilized EHR in multiple secondary contexts; however, there is a gap in developing an integrated and comprehensive framework that addresses the use of EHR in a standardized way for public health, privacy issue by anonymizing patient specific information, fusing multiple records with slight changes in the same information, augmenting a broad spectrum of contextual data, and so on. We present a framework that can be used in the context for acquisition and transmission of EHR from multiple sources as an evidence base for addressing public health-related activities, including surveillance, registries, and immunization record keeping while addressing all the gaps we have identified in the literature that are critical for developing countries. In addition, EHR data are also effectively processed to serve as a knowledge base for building artificial intelligence-based research models. We, in our model, utilize Health Level Seven (HL7) as an interoperability health standard and recommend creation of specialized data marts to support public health and research-related knowledge bases. The proposed framework in its adoption provides a very effective platform for generating alerts and alarms along with providing statistics for better planning of healthcare-related issues at national, district, or at any level of administrative hierarchy. It is applicable to any country even when there is no standard EHR and has hospitals working in silos with limited digitalization. We have validated this framework for its mapping to a national level public health hierarchy in Pakistan.

INDEX TERMS Electronic health records (EHR), evidence-based policy, health data interoperability, health information exchange (HIE), Health Level 7 (HL7), public health framework, public health surveillance.

I. INTRODUCTION

The public health policies and their impact on population relies on the evidence base collected for the decision making process. The evidence base is built using multiple parameters including population general characteristics as well as surveyed health statistics. Public health decision makers utilize surveillance as a tool to provide significant evidence base for action. However, in addition to surveillance, research and epidemiology also serve as vehicle to acquire the evidence base. These sources are interconnected and each may be conducted to inform the other. Most recently, Electronic Health Records (EHR) or Electronic Medical Records (EMR) are emerging as a strong tool not only for the primary intended use of cataloguing patients demographic and medical data but also for evidence based decision making at clinical level [1]. Therefore, it is only obvious to use EHR as a source for creating an evidence base in public health. The presence of longitudinal health data provides an opportunity to populate the traditional surveillance reports at population level [2]. While the secondary use of EHR has been thoroughly investigated and experimented with in the last decade, at the same time, significant achievements have been done by the research community in developing efficient deep learning algorithms [3]. The discretly coded data can be developed from the information present in EHR that is used to represent medical information based on a standard specification. This has inspired multi domain analytical techniques to be developed [4]–[7]. The collective use of these techniques needs to be addressed under a standard framework in order

The associate editor coordinating the review of this manuscript and approving it for publication was Xin Luo.

to achieve the full potential of secondary use of EHR including utilization of EHR as a population health representation source and performing Machine Learning(ML) and Artificial Intelligence(AI) based analysis to achieve the meaningful use of health data in public health domain.

Currently in many countries including Pakistan, public health related activities are done under multiple programs that provide highly controlled data for building intelligence over a population health and addressing policies for ongoing and emerging health related issues [8]–[10]. These programs deal with specific diseases, outbreaks or general well being of a population and may run in parallel inter-dependently or separately. However, data collected and transmitted is highly fragmented, duplicated among various programs and not stored as a central repository [11]. More importantly, the format of recording information varies from program to program and accuracy of data recorded is not verifiable [12]. There is also a lack of regulations to include health data from sources other than the government health care facilities.

In this article, we propose an improved framework called Public Health Framework (PHF) for public health analysis that can be realized through creating a pathway for national level regulations for collection and dissemination of health data through a collaborative framework that spans multiple governmental levels. At the base of this pathway, EHR can serve as source of electronically stored coded data to enhance public health decision making process. In addition, the presence of annotated data in EHR can be effectively utilized to create AI and ML models by researchers and made available for creating specialized analytical system for public health analysis and decision making process. This article describes PHF as standard based interoperability framework for creation of a pathway for acquiring and transmitting public health indicators from EHR evidence base to national level.

Therefore, unique from previously published work, PHF makes the following contributions integrated as a single pathway for health monitoring, informatics and research;

1) It allows the inclusion of clinically objective data from EHR into public health surveillance.
2) It allows representation of population visiting private health facilities by facilitating minimal configuration requirements and support of legacy systems.
3) It ensures availability of annotated data to create and train Machine Learning (ML) and Artificial Intelligence (AI) models
4) It provides a platform for creating publishable ML oriented models based on real time data.
5) It allows inclusion of context information over a population such as including social and economical determinants of health to perform multi dimensional analysis of health data in public health perspective.
6) It allows the possibility of populating public health surveillance reports with electronically sent EHR data as well as using EHR data in parallel with surveillance data.

The rest of this article is organized as follows: Section II describes the work done related to using EHR for secondary usage including public health informatics. It also gives a comparison of using EHR vs. surveillance techniques for public health. Section III describes the programs and their respective tools for population health data acquisition in Pakistan and their use at multiple health care levels. The section details Pakistan health care delivery system as a case study for proposed framework implementation. Section IV describes, in detail, the contribution of realization of the proposed architecture in public health and research context, the architectural layers (IV-A) and their corresponding components (IV-B). Section V discusses the objectives of population health improvement that this framework is able to address through realization of discussed architecture.

II. RELATED WORK

Owing to the large volumes of health data available through multiple sources, novel approaches to use big data for meaningful analysis and predictive patterns are being designed and implemented [13]. The medical standards and interoperability standards is now unified under health information exchange systems (HIE) for transmission of health data to public health authorities. EHR has been extensively proven useful for improving individual health. It has also emerged as a useful data source for population health analysis and aids in decision-making at the policy development level [14], [15]. Health Information Exchange (HIE) systems and frameworks are designed to provide secure, reliable and seamless exchange of health data among diverse systems and provide a protocol for health information flow with the objective of health care delivery services to population ranges. Therefore, Health data entities are moving forward from one to one data exchange towards interoperability standards for transmitting health data.

HL7 [16] is a set of standard, formats and definitions for exchanging electronic health records (EHR) and is accredited by American National Standards Institute (ANSI). HL7 standard defines a standard format that allows messages to be transmitted between disparate systems. It recognizes that all health care facilities are not identical and there cannot be a single standard protocol model to represent the data exchange process between these facilities. Therefore, it provides a broad range of messaging standards that can be applied to large scale health facilities, major hospitals as well as to stand alone diagnostic centres, laboratories, and clinics. It is now a widely accepted e-Health standard and is considered efficient for health information exchange. The HL7 v2 offers messages models, documents and services that cover the health data integration, retrieval and exchange requirements of most of the public health agencies. The work presented in this article is based on HL7 v2 standard specification for health information exchange. The details of the message generation and processes to follow are described in section IV-B.

Multiple successful systems are deployed world wide working from local to national administrative levels to

TABLE 1. A comparison of state of the art disease surveillance systems based on features common to surveillance systems in general. The table shows that the PHF system presented offers rich EHR attributes as source and more contextual data for meaningful multi dimensional analytics.

	Data Source	Indicators	Format	Transmission	HL7	Privacy	Context
EARS	Emergency Dept.	Chief complaints Admission and discharge codes physician office data	SAS datasets, Microsoft Access database tables, Microsoft Excel worksheets, delimited textfiles	Manual uploading	N	N	Schools, business attendance Drug sales 911 calls
ESSENCE	Clinical data, Nonclinical syndromic data Health events-related information.	Chief complaint	Text Strings for Chief complaints ICD 9 Codes	FTP Protocol	N	Disclosure control Data sharing policy	OTC Sales of pharmaceuticals Nurse hotline calls school absenteeism veterinary reports
BioPortal	Human and animal diseases	Chief complaint	Text Strings for Chief complaints	Batch Transmission	Y	Role Based	Dead bird sightings Mosquito control Information
RODS	EHR Registration	Age Gender Zipcode date/time of admission Chief complaint	HL7 specification	Real time	Y	Disclosure control	Not included
BioSense	Department of Defense (DoD) Military Treatment Facilities (MTF) Department of Veterans Affairs (VA) Laboratory Response Network (LRN) Electronic Laboratory Results (ELR)	Patient chief complaint Diagnosis, Patient demographics Hospital census ED-specific clinical data Microbiology test orders and results Radiology orders and results Medication orders	HL7 specification	Realtime Batch mode	Y	Role Based	Lab results from environmental sensors
PHF	Clinical EHR data Non Clinical data surveillance	HL7 attributes in MSGs: ADT ORM ORU (ADT types attributes outlines in Table 3)	HL7 Specification	Realtime Batch mode	Y	HIPAA Privacy compliant algorithm Role based Access	Education data Weather data Urbanization data Geographical data Population census data Water and sanitation data Pharmaceutical Sales Data Economic survey data

perform analysis on various public health diseases. The systems vary in data collection, transmission and analysis scale and techniques. Table.1 compares some of the most relevant existing state of the art systems for disease surveillance and also evaluates the presented framework PHF on the same parameters. The comparative analysis shows that most of the systems lack inclusion of context in analytical processing of data. The privacy protocol identifies, if there is any explicit provision of HIPAA compliance or how duplicate detection is employed for multiple records when HIPAA compliant attributes are omitted from the data. 'N' shows no explicit handling of HIPAA compliant attributes is done for privacy preservation.

The Early Aberration Reporting System (EARS) has been used for large scale event monitoring. Data are acquired from patient encounters from participating health care facilities. However, no standardized data reporting or transmission protocol is used.

Electronic Surveillance System for the Early Notification of Community based Epidemics (ESSENCE) [17] collects data from multiple sources including chief complaint ICD-9 codes, public data sources and other surveillance outcomes data. Therefore, the data is coming from multiple heterogeneous sources. However, the acquisition of multi sourced data introduces time lag in the system. While all the data is collected electronically, the transmission is not based on HL7 specifications that reduces the meaningful syndromic surveillance analysis. Essence provides a multi layered data dissemination approach that is analogous to our data mart approach. The layers use role based access for data distribution to multiple users.

BioPortal [18] provides analytical capabilities through web based portal based on monitored disease incidence data acquired from free-text chief complaints. 'P' in HL7 attribute in 1 indicates that system partially supports HL7 based inputs, however the transmission is not based on HL7 messages. Limited contextual data is also made part of the BioPortal such as surveillance data on dead bird sightings, mosquito control information, epidemiology data and genes sequence data, however, the number of indicators from EHR systems is very limited. Use of BioPortal in collaborative works has given meaningful results or disease specific studies [19].

Real-time Outbreak and Disease Surveillance (RODS) [20] acquires data through HL7 protocols including age, gender, home zip code, date/time of admission, and a free text chief complaint of the patient. The system provides powerful analytical capabilities, however, the EMR data sources are limited and there is no provision of incorporating contextual data for in-system multidimensional data analysis. Successful collaborative studies have been conducted [21].

BioSense [22] system employs HL7 messaging protocols to send multiple indicators including patient chief complaint, physician diagnosis, supporting patient demographic data, daily hospital census, ED-specific clinical data, microbiology test orders and results, radiology orders and results, and medication orders. Data transmission is supported in real time as well as batch mode. The system provides statistical reporting, time series and spatial graphs for interpretation by public health expert analyst. However, while the collaborative studies are frequently conducted with the data collected through BioSense system [23], the architecture does not integrate in itself the other contextual resources for multi dimensional analysis of data.

While the systems described above offer powerful analytical state of the art capabilities for health informatics based on validated machine learning algorithms, the data collection mechanism and indicators being collected may limit the outcome due to absence of useful EHR data attributes of public health interest. A more detailed comparative study conducted on surveillance systems adopted world wide has shown that the under developed countries, while requiring most of these public health interventions, lack the infrastructure for these

TABLE 2. **A comparison of properties of disease surveillance systems based on literature study when EHR and traditional surveillance sources are used. The table shows that using EHR as source as compared to other reporting tools provide a clear advantage.**

	EHR Based Surveillance	**Traditional Surveillance**
cost [24]	Low cost when automated processes are already in place at health care facilities recording patients information.	higher cost due to parallel programs of data collection + cost of data pre processing and annotation
data accuracy [25]	Dependent on existing data acquisition processes. Can be ascertained for a subset of data	Strong quality control for pre defined data elements
timeliness [26]–[28]	realtime data acquisition and analysis possible	mostly historic data available, time period defined by program
completeness [24], [29]	lower due to recording errors and omissions	higher due to required fields compliance requirements
data range [30], [31]	broad in terms of population, can be increased or decreased based on required study sample	depends on sample collection frequency and population set defined by study
pharma and medication data [32]	detailed and recent data available	details available pertaining to disease or focus subjects available over a certain period of time
indicators [13]	large number available that can be selected as per program requirement and filtered accordingly	predefined derived, aggregated or direct indicators available useful for a certain analysis type
result reporting [33]	multiple possibilities of reporting formats, dimensions, and tools for report analysis	limited by self reporting nature, statistical measures and cross sectional reports
time series analysis [34]	possible because of availability of time series data	not possible because of data in the form of cross sectional reports

state of the art systems [35]. In addition, the timeliness and completeness of EHR data transmitted through a standard based protocol can overcome the limitations in the afore mentioned architectures. Therefore, placing EHR at the base of these frameworks as evidence base can potentially improve the analytical algorithm performance at the upper layers. The design of EHRs implicitly includes a standardized measure of disease and a significant representation of well being and health status of a definable population. While EHR systems are constantly being improved to represent real time longitudinal health data of a population, it is important to note that in order to create meaningful and Artificial Intelligence (AI) based models for public health analysis, the availability of annotated data is of utmost significance. This data can be made available through a standard based architecture that overcomes the existing challenges present due to self informed nature of data available through traditional disease surveillance systems [2], [36]. The comparison of using EHR data for populating surveillance reports and using traditional surveillance means have been studied and analyzed in literature. Table.2 gives a general comparison of the properties of an effective population disease surveillance system when EHR data or traditional surveillance means are used. Additionally, there is a difference between the information required from EHR for individual healthcare and for public health care [37]. To provide information about populations rather than individual patients, it is important to define a real time, accessible and standardized way of transporting the samples of population heath status to concerned entities such as AI practitioners, government analysts and decision makers [38]. Therefore, such a framework must, at its fundamental level, be able to identify and defines the intersecting health indicators at individual and population level as well as consider the inclusion of EHR data into public health administrative data in parallel with surveillance data [39]. The primary goal of EHR/EMR systems is to organize, store and categorize patients' records. The challenge arises when this designed-for-storage data has to be retrieved for transmission or for sharing with other systems. Significant research has been conducted to design frameworks focused on information sharing among multiple EHR systems [1], [40]. The secondary use of EHR has also been investigated for public health interest [41].

The existing frameworks also have limited contextual sources such as urbanization, socio economic and education census, sanitary conditions etc. These factors play an important role in under developed countries where environmental and other factors are constantly changing and evolving. In addition, the annotated datasets can be made available to researchers after anonymization through domain specific data marts.

Our presented framework (PHF) shows the implementation details on the instance of Pakistan health care delivery system with low technical resources such as HL7 compliant EHR systems. Furthermore, the components of presented framework can be open sourced to be re used under different domains other than human disease outbreaks, such as agriculture and wildlife diseases etc. Therefore, to the best of our knowledge no work thus far has given an integrated standard based approach for context based standardized acquisition, transmission and analysis of EHR based public health data while staying privacy compliant and also addressing the need of research community through annotated, anonymized datasets. Therefore, there is a need to develop a framework model that can then be used to inform decision-making by projecting the potential outcomes associated with different policy decision. In addition, the presented framework presents a widely accessible central repository of labeled data to AI practitioners in the domain.

Thus, while recognizing the effectiveness of using EHR as evidence base for population health improvement, our public health standard HL7 based infrastructure framework serves the purpose of information exchange and maintenance of a central database as a national health repository. The objective of this system is to provide reliable, standardized, consistent and up-to-date health information to health ministry, health care authorities, and other administrative levels of health care delivery system in any country where limited digitization hinders the meaningful analysis and effective outcomes in public health sector. It also enables the creation and access of labeled data sets for research in a particular domain.

III. STATE OF PUBLIC HEALTH TOOLS IN PAKISTAN

In order to test the validity of the presented framework, we have applied it to the health care system in Pakistan that has helped in identifying challenges and steps involved in real time implementation of framework. It is important to understand how in this context, the public health tools are currently being practised for informed decision making.

The health care delivery system in Pakistan consists of a complex interconnected and interdependent, governmental and non governmental facilities. It can be mainly categorized into private, public and non government or non profit sectors. The public sector consists of federal and provincial governments where federal governments looks upon research institutes, ministry of Defence, ministry of health, ministry of Inter Provincial Coordination (IPC), Ministry of Health Service Regulation & Coordination (HSRC) in addition to other ministries and provincial governments are responsible for various departments including provincial health departments.

Fig. 1 shows the interconnected health care delivery system hierarchy in Pakistan with Health Expenditure (HE) in each major sector. The public health sector includes military health care systems, cantonment boards' health departments and provincial government health care systems. The federal government is also responsible for running vertical health programs through primary to tertiary levels. The territorial government health systems are managed at district level. The district level consists of primary and secondary health levels that are implanted through Basic Health Units (BHUs), Rural Health Centers (RHCs), Maternal and Child Health Centers (MCHCs), Dispensaries and Tehsil Headquarter Hospitals (THQs), District Headquarter (DHQ) Hospitals respectively. The tertiary health level is implemented through health care facilities that are handled directly by the provincial health departments and the federal government.

The Health Management Information System (HMIS) was developed in Pakistan in 2001 under "health for all" approach and implemented at all three levels, primary, secondary and tertiary [42]. It is designed to collect and transmit information from BHUs and RHUs related to service delivery and utilization, logistics and disease surveillance to district HIS at DHQs. Districts are empowered to self contain and record and transmit population and health statistics to federal level through respective provinces [43]. It has since then gone under various changes and improvements under multiple health improvement frameworks and programs developed and implemented over the years [44].

Both passive and active surveillance programs are conducted that provide general health trends and emerging diseases or disease outbreaks respectively. Multiple vertical public health programs run independently to each other using different HIS tools such as District Health Information Systems (DHIS), Lady Health workers Management Information System (LHWs MIS), Maternal and Child Health(MNCH), HIV/AIDS, TB, Expanded Program on Immunization(EPI), Dengue and Malaria Control MIS, Service Statistics, Logistics Management Information System (LMIS), Financial & Human Resource MISs, Disease Early Warning Systems(DEWS), Notifiable Disease Surveillance(NDS) and several surveillance databases and dashboards operating according to their specific needs and respective capacities across all provinces and at the federal level [45]. However, the data collected from these facilities is program specific and in many cases redundant. In order to create a complete population health story, all annotated EHR data must be included and not just the summary reports of incidences of disease occurrences. Most of the data recorded is entered by the particular programs dedicated focal person at the heath facility. The collected data is consolidated and transmitted to district level as per the programs guidelines.

Another important aspect of the vertical public health programs is the absence of data from sources other than the public sector health care facilities. Due to inadequate health regulations in the private sector, only a small percentage of private sector health facilities participate in the public national health programs. According to National Health Accounts (NHA) report for fiscal year 2015-2016 [46], 63.4% of total health expenditure is funded by private health set ups, 89% of that is done by the private households referred to as out of pocket expenditure. This implies that a vast majority of the population is visiting private health facilities but the data produced is not included fully in the health programs thereby compromising the analysis outcomes of these programs.

According to National Health Vision (NHV) 2016 – 2025 [47] of Pakistan under Ministry of National Health Services, Regulation, and Coordination, a need for central integrated repository for evidence-based decision making, policy formulation, and health systems research has been identified. According to the NHV report, the current surveillance systems require analytical capabilities with reference to local challenges, health budget and systems' capacity. While considerable progress has been made in the past decade, national surveillance does not yet fully reflect the recommendations from NHV to make systemic improvements.

In 2018 Prime Minister Task Force On IT & Telecom was formed to work on data standards and annotations for their consolidation for planning improvements in health care delivery to masses. This work is part of the program, proposing a framework that can be adopted for this national initiative.

IV. PUBLIC HEALTH FRAMEWORK (PHF)—AN EHR BASED HEALTH DATA CONSOLIDATION FRAMEWORK

We present a comprehensive and integrated framework consisting of a layered architecture with components placed strategically to allow seamless and privacy compliant

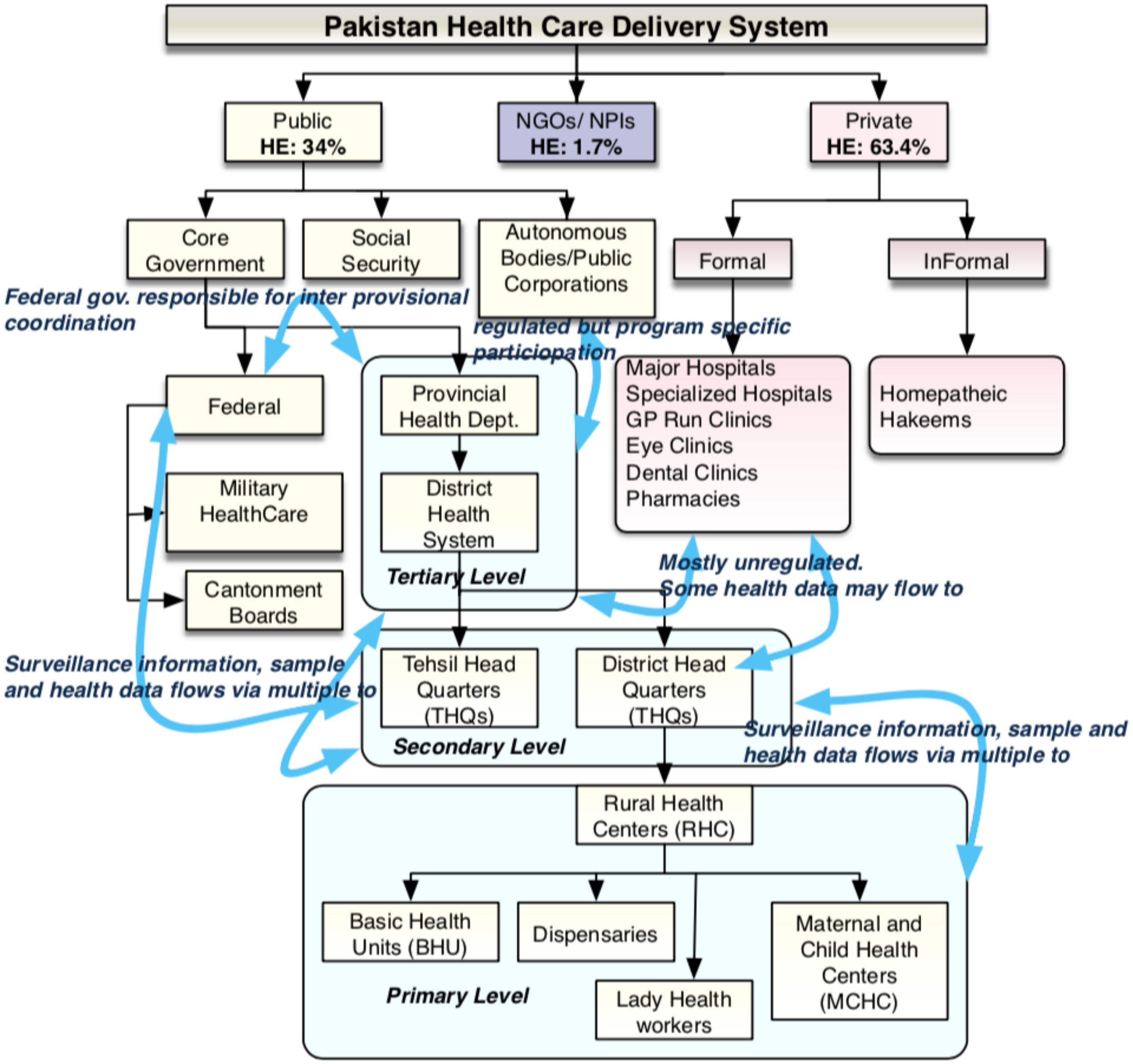

Pakistan Health Care Delivery System
Public HE: 34%
NGOs/ NPIs HE: 1.7%
Private HE: 63.4%
Core Government
Social Security
Autonomous Bodies/Public Corporations
Formal
InFormal
Federal gov. responsible for inter provisional coordination
regulated but program specific participation
Federal
Military HealthCare
Cantonment Boards
Provincial Health Dept.
District Health System
Tertiary Level
Major Hospitals
Specialized Hospitals
GP Run Clinics
Eye Clinics
Dental Clinics
Pharmacies
Homepatheic
Hakeems
Mostly unregulated.
Some health data may flow to
Surveillance information, sample and health data flows via multiple to
Tehsil Head Quarters (THQs)
District Head Quarters (THQs)
Secondary Level
Surveillance information, sample and health data flows via multiple to
Rural Health Centers (RHC)
Basic Health Units (BHU)
Dispensaries
Lady Health workers
Maternal and Child Health Centers (MCHC)
Primary Level

FIGURE 1. The health care delivery system of Pakistan is a hierarchal interconnected system with Health Expenditure HE being highest in the private sector. The flow of health information among entities under different public health programs is shown. The data collected at primary and secondary level flows to tertiary level where its is compiled and submitted to federal government for report compilation. Only a part of private sector is represented by these public health programs.

acquisition and transmission of EHR data of public health interest at all levels of health care delivery system. Fig.2 presents the layers and their respective components in the framework. In PHF different hospitals and other data sources communicate with the central server using a secured communication channel and a communication protocol. Developing this framework can help to identify the disease outbreak in its epidemiology context using multiple parameters and evidences. The model can then be used to inform decision-making by projecting the potential outcomes associated with different policy decisions.

A. CONCEPTUAL FRAMEWORK VIEW

This section presents the different layers in PHF and their role in acquisition, transmission and storage of EHR data. As represented in Fig.2, the bottom most layer is the glue layer that resides at the source site. Middle layer is the resource fusion layer, that stores the public health data and ensures is anonymized and appropriate availability to different entities. The upper most layer is the service layer where data is accessed from middle layer and contains the business logic and intelligence.

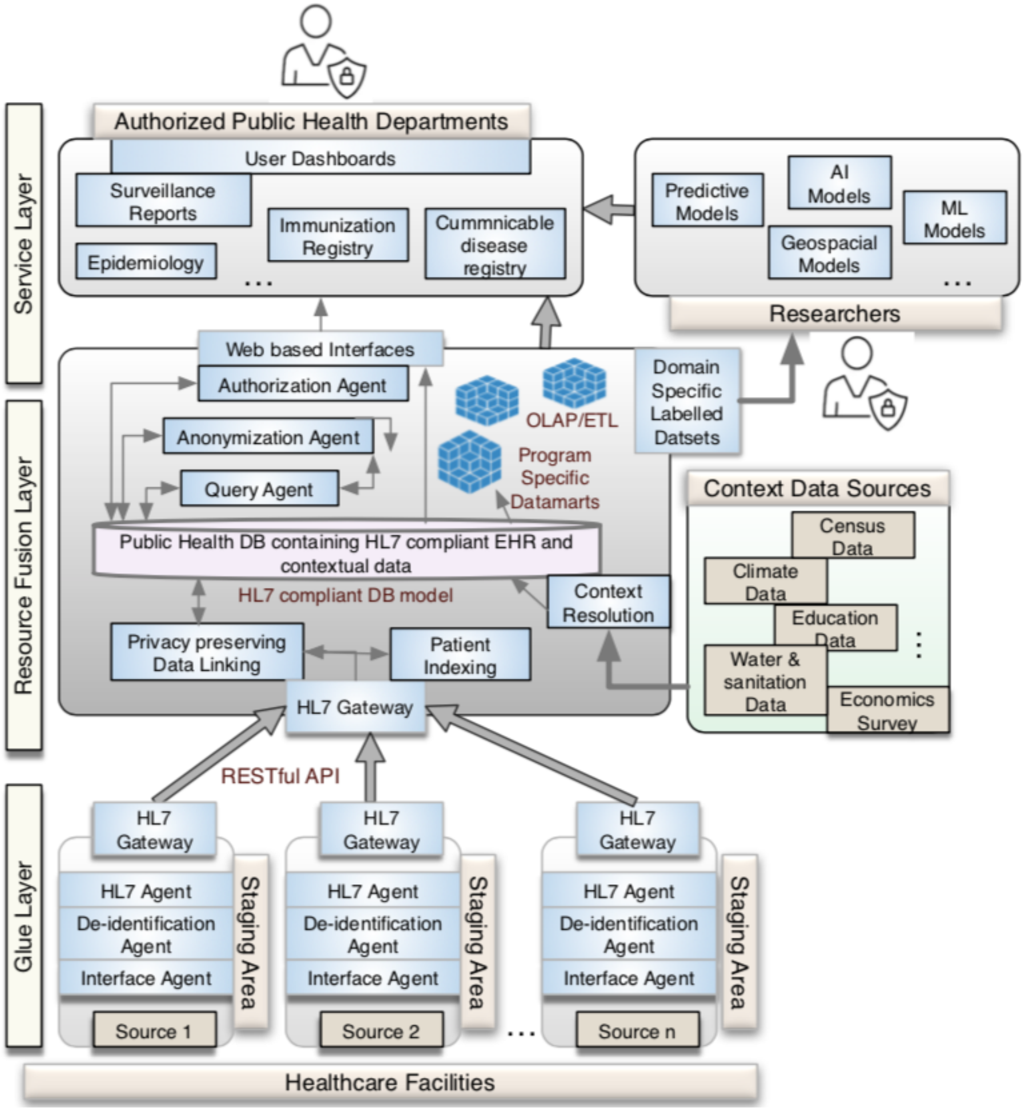


FIGURE 2. **The overview of proposed layered national framework for public health informatics and research modeling. The glue layer is implemented at primary and secondary level as well as in private hospitals participating in the program through adequate regulations. The resource fusion layer and service layer are implemented at tertiary and federal level of Pakistan health care delivery system.**

1) GLUE LAYER

In order to communicate information from different EHR systems to a public health databank, an agreed upon communication and document standard has to be used. However, forcing all HIS to use same standard and protocol for health data communication is infeasible and impractical. Therefore, we create an intermediate glue layer that acts to mediate the transmission of health data from heterogeneous legacy EHR systems to the population databank. This allows EHR system of hospitals to communicate with the HL7 agent independent of the types of standards being followed internally as well as by other EHRS. This also implies, that for every different HIS system, only a new glue layers needs to be implemented and other components of the system are able to communicate with the HIS through this glue layer. The mapping of EHR features to HL7 features is done based on use case modeling of the source system. The glue layer queries the legacy system to extract public health indicators and converts them into hl7 format. The glue layer primarily performs two operations; mapping and formatting. Mapping is a translation of EHR attributes to the HL7 attribute values where as formatting involves converting EHR attributes to HL7 compliant types and attribute values to a standard vocabulary terms and create an HL7 message from the required patient data. The mapping is done through creating a configuration that maps the EHR attributes to HL7 fields. The HL7 agent at glue layer is responsible for patient de-identification and creating an HL7 message ready to be transmitted to the population databank server through an HL7 gateway. Thus, this layer preserves the privacy of patient data. This layer also hosts a staging area that is used to prepare, process and transform the data received from the source to be transmitted through the HL7 gateway.

2) RESOURCE FUSION LAYER

The resource fusion layer consists of multi dimensional components covering authorization of users, query processing, integration, linking and storage of incoming HL7 messages as well as anonymization of data for creating publishable health datasets. When data is received from HL7 gateway, this layer is responsible for integrating incoming anonymized data with the existing records in population databank. For this purpose, a duplicate detection algorithm is employed that matches the hashed values of incoming data to the existing records and creates a decision tree to find matching records to fuse. It is possible to fuse data and achieve high quality linkage without the need to disclose fully identifying personal information. The master patient index implemented at this layer allows for synchronizing incoming EHR record with existing EHR data. Multiple other external sources are accommodated at this layer that constitutes the health context, such as geographical data, environmental data, education data, population census data and other social determinants of health etc. The transformation, transmission and storage design of the context data is not covered under the scope of this paper. This layer therefore hosts a data warehouse in the form of population databank. The presence of integrated, organized and categorically defined data in the databank illuminates multiple opportunities. Subject or study specific data marts containing pre-processed standardized data marts are spawned out from this databank for application or program specific purpose that can also serve as domain specific labeled datasets for research or as a program specific reporting tool, while on the other hand, the population databank can be queried directly for a population wide depth analysis of health and other data for policy and regulation decisions.

3) SERVICE LAYER

This layer constitutes the business analytics, custom public health applications, AI and ML based predictive models, implementation systems, report generation tools and public health dashboards based on the respective data marts spawned from underlying population databank. The outcomes generated from these models and systems can help public health authority in prioritizing public health resources and to support evidence based decision making. EHR can provide reliable evidence in addition to traditional surveillance systems for creating models based on AI and ML. Inclusion of AI and ML predictive modeling in public health can enhance real time analysis. The accuracy of these models previously being built has been affected due to unavailability of standard based real time labeled datasets having verifiable content. The availability of multi domain labeled data sets will allow creation of effective research models allowing the gap between research products generated and requirements of health care policy makers to be reduced.

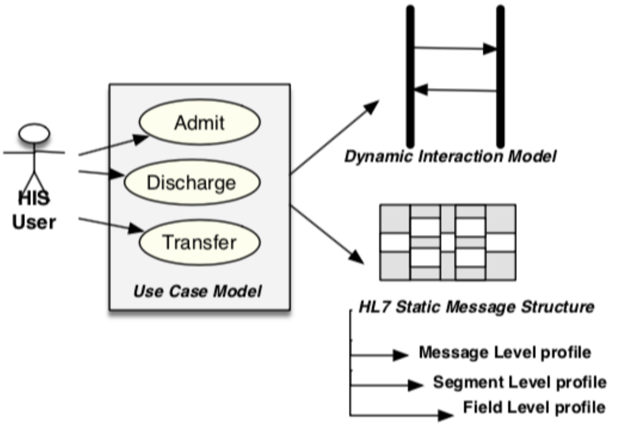


FIGURE 3. The overview of process of mapping source attributes to HL7 message mapping. The use case model of the source system is analyzed to create corresponding static HL7 message model consisting of message structures as well as dynamic interaction model consisting of activity or state diagrams.

B. INTERNAL VIEW

This section provides an introduction to the various components carried at different layers of PHF with a focus on glue layer components as this layer is a significant key player in framework scalability and interoperability.hl These components and their respective position in the layered architecture of PHF are presented in Fig.2.

1) INTERFACE AGENT

In order to interact with the data being generated and stored at the health care facilities, it is important to understand the clinical flow of data, the semantics of data, design of conceptual model and physical storage model of data at a particular site. The legacy systems are able to provide a data dictionary, documentation or workflow of the health care facility. An automated interface agent is only able to communicate with the system when the source to target mapping of the existing EHR system elements has already been studied and documented. This requires an understanding of the transactions taking place inside a health care facility using interviews, study of existing documentation and meta data. Once these transactions are understood, the interface agent can be automated to query attributes of interest from the existing EHR systems.

Therefore, use case derived HL7 message profiling is performed. All use cases are analyzed for an unambiguous specification of corresponding HL7 messages mapping. Fig.3 shows the source to HL7 compliant mapping process example where HL7 message profiles and interaction models are derived from the use case models of the source HIS. The use case derived mapping includes the understanding of source attributes semantics, data types and user defined groups. The interface agent must also be able to improve the data quality without compromising the integrity of data. The data processing done at this layer includes data type transformation from source to HL7 compliant types, for example date type conversion; data codes mapping from source to standard codes, for example, diagnosis codes from internal coding scheme to ICD 9-10 coding; data splitting from single source elements to multiple HL7 fields. This processing is done in the staging area at the glue layer. The staging area hosts the terminology server, code conversion tables and error log tables for preprocessing steps.

This concludes that given that the initial internal workflow of healthcare facility is studied and meanings of internal vocabulary of a facility are translated to a unified format, the interface agent; a) queries the relevant information from the database, b) perform data formatting, data type and content checking. c) Convert the values into a standard terminology and stores the data in the staging area.

2) DE-IDENTIFICATION AGENT

The HIPAA Privacy Rule is US national standard that requires that appropriate safeguards be applied while disclosing or exchanging patient health information. It identifies 18 patient identification attributes that must not be present in an identifiable form that could enable re identification of the patient. While we are not interested in personal identification of a patient in public health perspective, these attribute(s) are required to fuse data received from multiple sources. In order to make the dataset HIPAA compliant, the agents installed at the client site encode the patient identifying attributes before transmission. The agent applies SHA256 to all patient-identifying attributes. SHA256 is a cryptographic Hash Algorithm defined by National Institute of Standards and Technology(NIST). Since its not an encryption algorithm, it is not possible to decrypt the attributes to the original form. The SHA algorithm creates a unique hash for every data element. The typographical errors, phonetic errors and other typing mistakes must be first transformed using patient MR number or identification number or by using some encoding functions such as soundex or metaphone, whichever maybe applicable in the context. This allows the linking of hashed patient information more robust to the homonym errors occurring due to original data problems. The de-identification agent also assigns a new identity to the patient for representation in the public databank based on the source of the data and some of the hashed valued attributes. This id is synchronized with the master patient index at the resource server. It allows indexing of the patients records based on the attribute tags and facilitate querying while staying privacy compliant.

3) HL7 AGENT AND GATEWAY

The HL7 agent is responsible for creating an HL7 message from the anonymized hashed data as well as clinical and demographic attributes, ready to be transmitted to the public health databank server through an HL7 gateway. The HL7 mapping of the clinical data attributes and anonymized data attributes is done by creating a predefined transformation procedure. The mapping establishes a semantic equivalence between HL7 message constructs and data elements at the host facility.

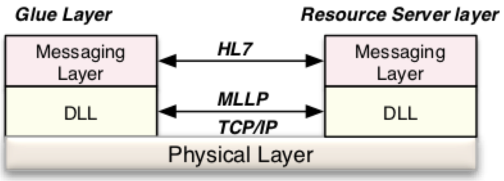
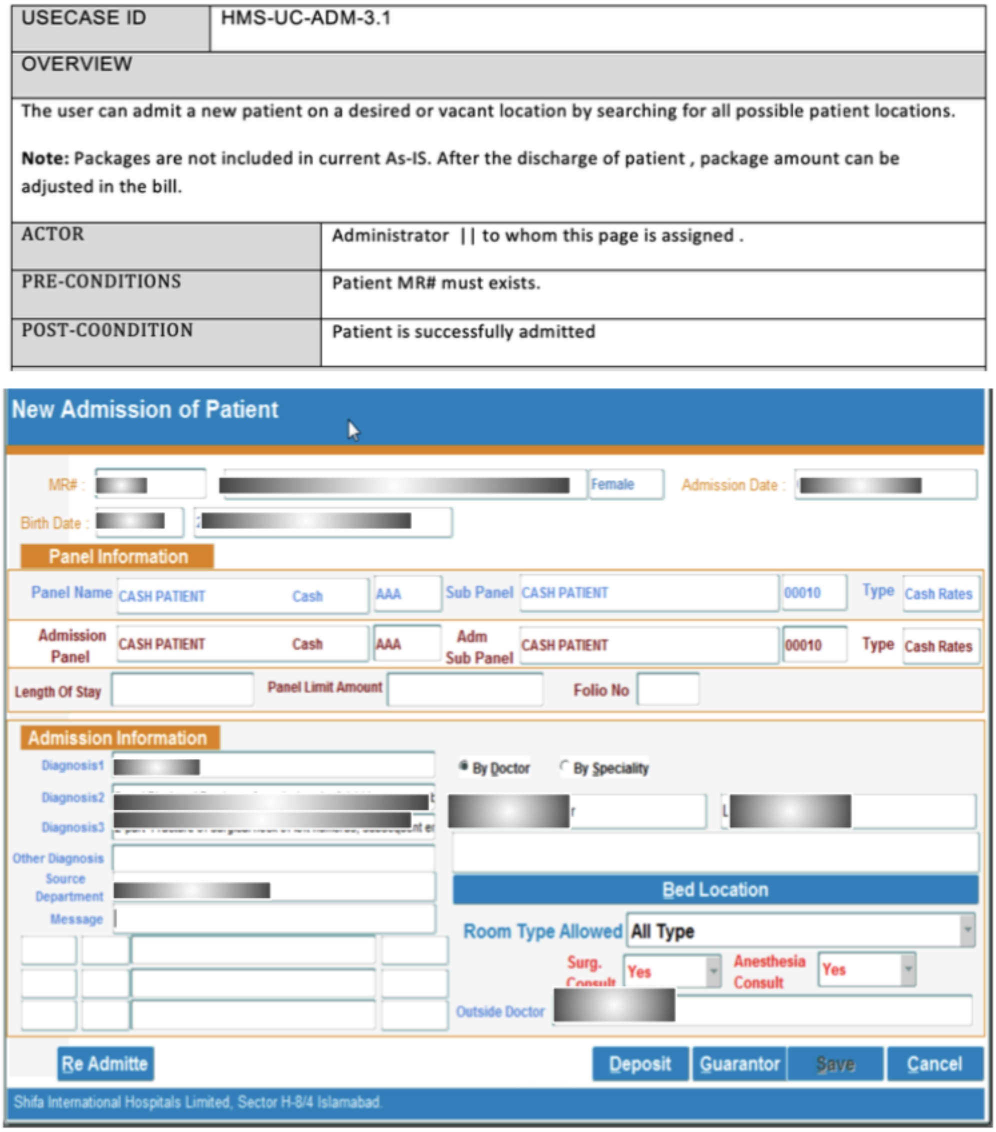


FIGURE 4. The multi layer HL7 gateway implementation architecture for exchanging HL7 messages consisting of physical layer, Data Link Layer DLL and message layer.

HL7 V2.x [48] gives an interoperability model for health data and is the most widely used standard. It consists of messages, that are composed of segments that are further categorized into fields where actual data is stored. Each message has a type and a trigger associated with it. There are more than 183 segments included in HL7 v2.x. When an event occurs such as patient admission, patient transfer, patient discharged etc., it defines a trigger and is represented by a particular type of message. Messages can have a single trigger and thus a single format (e.g. VXU message) or they can have multiple triggers and formats (e.g. ADT message). Segments in the messages can be optional, repetitive or both. The messages are populated only with the required data. HL7 v2 defines multiple message types with Admission, Discharge and Transfer (ADT), that includes the patient demographics, being the most commonly exchanged message.

Our HL7 Gateway interface accepts multiple HL7 message types for transmission using Transport Control Protocol/ Internet Protocol (TCP/IP) layers with Minimum Lower Layer Protocol (MLLP) as a wrapper protocol on top at data link layer Fig.4. It is responsible for connection establishment, transmission and connection termination between source and target. The physical layer transmits the binary data from source glue layer gateway to resource fusion layer gateway. The staging area serve as a repository for HL7 information to be drawn from the health care facility. Most of the error detection and correction during physical data transmission is handled by TCP/IP protocol and are not addressed through any supplementary code. Messaging layer use HL7 infrastructure for initiating message exchange between two gateways.

As an HL7 messaging layer example, we present a patient admission use case for a private hospital, Shifa International Hospital (SIH) in Islamabad, Pakistan. We create a mapping of patient information which is derived from workflow model and use case model. Fig. 5 represents the use case for 'Admit a Patient' and the associated graphical user interface. While the mapping is generated from the logical data model, the user interface gives the understanding of data components collected when a patient is admitted through the system. Most of the attributes are directly mapped onto HL7 fields, while others are extracted through the links in logical data model. The patient admission use case is represented by ADT (Admission, Discharge, Transfer) message in HL7. The intent of ADT messages in HL7 is to carry patient information including patient demographics, registration, admission, discharge



USECASE ID	HMS-UC-ADM-3.1
OVERVIEW	
The user can admit a new patient on a desired or vacant location by searching for all possible patient locations. **Note:** Packages are not included in current As-IS. After the discharge of patient , package amount can be adjusted in the bill.	
ACTOR	Administrator || to whom this page is assigned .
PRE-CONDITIONS	Patient MR# must exists.
POST-COONDITION	Patient is successfully admitted

New Admission of Patient

MR# : [redacted] [redacted] Female Admission Date : [redacted]

Birth Date : [redacted] [redacted]

Panel Information

Panel Name	CASH PATIENT Cash	AAA	Sub Panel	CASH PATIENT	00010	Type	Cash Rates
Admission Panel	CASH PATIENT Cash	AAA	Adm Sub Panel	CASH PATIENT	00010	Type	Cash Rates

Length Of Stay Panel Limit Amount Folio No

Admission Information

Diagnosis1 [redacted]

Diagnosis2 [redacted]

Diagnosis3 [redacted]

Other Diagnosis

Source Department [redacted]

Message

● By Doctor ○ By Speciality

[redacted] [redacted]

Outside Doctor [redacted]

Bed Location

Room Type Allowed All Type

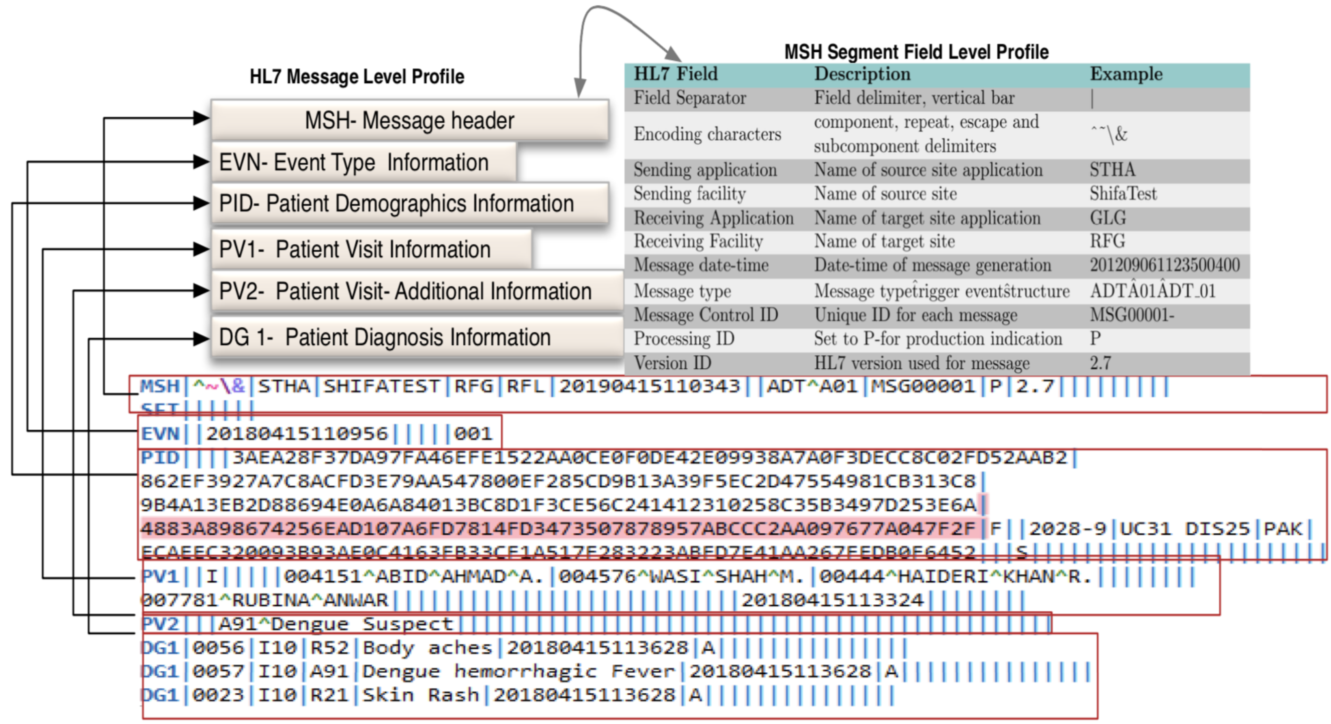
Surg. Consult Yes Anesthesia Consult Yes

Shifa International Hospitals Limited, Sector H-8/4 Islamabad.

FIGURE 5. The test site use case 'Admit a Patient' and graphical user interface representing fields for underlying data model.

and transfer etc. There are 51 different types of ADT message triggers that carry patient and patient related event communication information. As a significant example of using HL7 message and gateway to transmit data from a health care facility to public resource server, we used HL7 Application Programming Interface (HAPI), an open source java based library that implements HL7 message specification, in an Eclipse IDE environment. For the purpose of this document, we present an implementation use case of ADT 01 message that defines the patient admission. The particular field profile that needs to be transmitted via HL7 gateway to resource fusion layer depends on the indicators identified as being required in public health perspective and how the source EMR system manages this information. Not all EMR attributes are needed at the resource fusion layer for population representation. Therefore, while following the use case based HL7 mapping, each required attribute in the source EMR must be carefully placed to the corresponding HL7 field.

For ADT01 message transmission at SIH test site, an HL7 message is created. Contents for the segment fields are extracted from two tables namely, Patient and Admissions that contain data about operational transactions regarding patient admission and admission details respectively. Table3 describes the indicators extracted from the tables and the corresponding field level profile for ADT 01 message generated. The Message header (MSH) segment, Event Information (EVN) segment, Patient Information (PID) segment, Patient Visit (PV1) segment, Patient Visit-Additional Information (PV2) segment and Patient Diagnosis (DG1) segment are created as shown in Fig.6. The Message Header (MSH)


HL7 Message Level Profile
MSH- Message header
EVN- Event Type Information
PID- Patient Demographics Information
PV1- Patient Visit Information
PV2- Patient Visit- Additional Information
DG 1- Patient Diagnosis Information

MSH Segment Field Level Profile

HL7 Field	Description	Example
Field Separator	Field delimiter, vertical bar	|
Encoding characters	component, repeat, escape and subcomponent delimiters	^~\&
Sending application	Name of source site application	STHA
Sending facility	Name of source site	ShifaTest
Receiving Application	Name of target site application	GLG
Receiving Facility	Name of target site	RFG
Message date-time	Date-time of message generation	201209061123500400
Message type	Message typeˆtrigger eventˆstructure	ADTÂ01ÂDT_01
Message Control ID	Unique ID for each message	MSG00001-
Processing ID	Set to P-for production indication	P
Version ID	HL7 version used for message	2.7

```
MSH|^~\&|STHA|SHIFATEST|RFG|RFL|20190415110343||ADT^A01|MSG00001|P|2.7|||||||||
SFT||||||
EVN||20180415110956||||001
PID||||3AEA28F37DA97FA46EFE1522AA0CE0F0DE42E09938A7A0F3DECC8C02FD52AAB2|
862EF3927A7C8ACFD3E79AA547800EF285CD9B13A39F5EC2D47554981CB313C8|
9B4A13EB2D88694E0A6A84013BC8D1F3CE56C241412310258C35B3497D253E6A|
4883A898674256EAD107A6FD7814FD3473507878957ABCCC2AA097677A047F2F|F||2028-9|UC31 DIS25|PAK|
ECAEEC320093B93AE0C4163FB33CF1A517F283223ABFD7E41AA267FEDB0F6452|||S||||||||||||||||||||||
PV1||I|||||004151^ABID^AHMAD^A.|004576^WASI^SHAH^M.|00444^HAIDERI^KHAN^R.||||||||
007781^RUBINA^ANWAR||||||||||||||||||||||||||||20180415113324||||||||
PV2|||A91^Dengue Suspect|||||||||||||||||||||||||||||||||||||||||||
DG1|0056|I10|R52|Body aches|20180415113628|A||||||||||||||
DG1|0057|I10|A91|Dengue hemorrhagic Fever|20180415113628|A||||||||||||||
DG1|0023|I10|R21|Skin Rash|20180415113628|A||||||||||||||
```

Example ADT A01 Message

FIGURE 6. **The message level profile for ADT message and field level profile for MSH segment of the message is shown. An example ADT A01 message generated through use case profiling for test hospital is also shown. The highlighted part of the message reflects the data sent in the hash code form in order to perform privacy preserving linkage of records.**

segment contains the message type, trigger event, sending and receiving facility, and application details. This segment is included in all HL7 messages. An HL7 message is a sequence of all required and optional segments defined for a message type. For example, our example ADT message has 6 segments where MSH, EVN, PID and PV1 are the required segments that cannot be omitted as per the v2 specification where as PV2 and DG1 are optional segments that are included to represent information defined at our test site. The segments combine to constitute the message level profile for A01 message. The segments carry the data in fields that have a particular sequence. The mapping of test site data to the corresponding fields is represented through their position in the segment fields as shown in Table 3 that shows the field level profile of each segment.

This message is created and sent to HL7 ADT gateway for transmission. The ADT gateway follows a transmission protocol where an ADT message is initiated and sent by the SIH HL7 agent via the glue layer gateway. Therefore, this interaction follows a push based approach as shown in Fig.7a. Each received message received at the resource fusion gateway is validated by examining the MSH segment's 3 fields, that is; message type field (MSH.9), processing ID field (MSH.11), and version ID field (MSH.12). These fields must have a content that is acceptable to the application at resource fusion gateway. For example, ADT01 message will not be accepted by ORU message listener at the gateway. Similarly if the gateway accepts version 2.7 of the HL7 messages then other version messages will be rejected. If message received by the resource fusion layer ADT gateway is acceptable based on MSH segment, it is acknowledged by sending an HL7 ACK message that contains an MSH and an MSA segment as shown in Fig.7b. The MSA.1 field defines the acknowledgement code that is set to AA (Application Accept) if the message is accepted and MSA.2 contains the ID of the sent message. If the protocol validation fails or application fails to process message and it times out, the MSA.1 is set to AR (Application Reject) and sent back to the source gateway. The timed out message can be resent depending on the resend limit set at he source gateway configuration settings.

4) MASTER PATIENT INDEX AND LINKING AGENT

The Master Patient Index is maintained at the resource server in order to uniquely identify each patient record. Note that, this index exists in order to correctly link the inter visits and multiple visits data coming from multiple sources and is not meant for identification of a patient in real world. Maintaining this index is significant for correct fusion of incoming data with the existing records in order to maintain longitudinal advantage of EHRs. However, keeping a master patient index is costly and have to maintain the privacy concerns of sensitive patient information. Many sophisticated techniques have been designed and developed for accurate record lining but are circumscribed because of privacy policies such

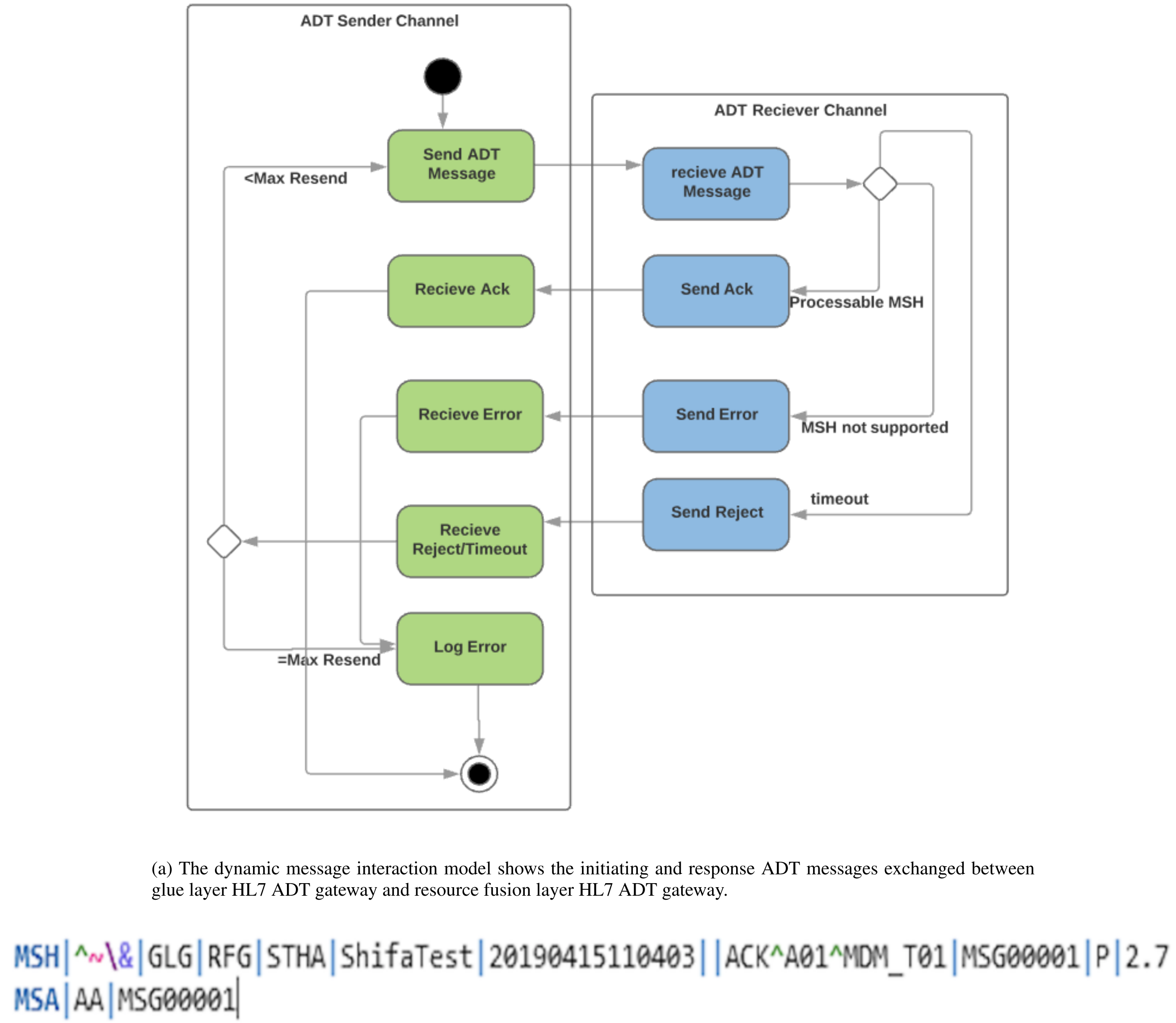
TABLE 3. Segment and field level profile of all required and optional segments except MSH segment of the ADT message generated using 'Admit a Patient' use case of a test site. C represents the cardinality of each mapping.

Source EHR element	Description at source	C	Implementation Guide	HL7 Field
EVN Segment				
Site ID	Unique identifier for the source EHR site where patient event occurs.	1..1	Every site is registered at the resource fusion layer via its glue layer.	EVN 7.2
Site Name	Name of the source EHR site	0..1	Stored during site registration process. May not be included in every message.	EVN 7.1
Reporting Date-Time	Date and time of event generation at source	1..1	Date and time remains as source at all layer levels.	EVN-2
PID Segment				
MR#	Patient internal Medical Record Number	1..1	Sent in hashed form for linking agent.	PID.4
Age	Patient age at the time of event	0..1	Not sent in message, as it is a derived attribute	-
Date of Birth	Patient date of birth.	0..1	Stored during patient registration event. Also part of ADT 01 message	PID.7
Gender	Patient gender	0..1	Stored during patient registration event. Also part of ADT01 message.	PID.8
Adress	Patient address	0..1	De-identified to union council or district level. Send original in hashed form.	PID.11
Contact	Patient contact number	0..1	Stored during registration event. Also part of ADT01 message Sent in hashed form for linking.	PID.13
PV1 Segment				
Patient Class	Identify class of patient	1..1	different from source department. Classify patients according to predefined categories .	PV1.2
Length of stay	number of days from admission	0..1	used for calculating discharge date of patient when not available.	PV1.45
DoctorID DoctorName Speciality ID	ID of already registered doctor Name of already registered doctor Specialty ID of already registered doctors from defined specialties	0..* 0..* 0..*	All three elements are stored for Referring doctor, Consulting Doctor, Admitting doctor	PV1.8/ PV1.9/ PV1.17
Source Department	department requesting admission	0..1	Used as hospital service used by Patient prior to admission	PV1.10
Admission Date	Admission date and time of Patient	1..1	Patient admission date time in format: YYYYMMDDHHMMSS	PV1.44
PV2 Segment				
Admission Reason	Text containing admit condition or diagnosis	0..1	ICD-9, ICD-10 codes of admission diagnosis, free text describing reason for admission	PV2.3
DG1 Segment				
Diagnosis 1 Diagnosis 2 Diagnosis 3	ICD-10 code of patient diagnosis ICD-10 code of patient diagnosis ICD-10 code of patient diagnosis	0..* 0..* 0..*	Multiple diagnosis sent. The first diagnosis is considered as primary diagnosis. For each diagnosis diagnosis type, coding method (ICD-9, ICD-10 code),diagnosis, description, and date time is included.	DG1.1- DG1.6

as HIPAA. Other techniques mostly commonly referred to as privacy preserving techniques are more appropriate in health case scenarios [49]. In our case, the master patient index maintains the identifying information with the help of some core data elements in their hashed forms that are used in the fusion algorithm by the linking agent. The irreversible nature of hashing technique used allows fusion of records without releasing the patient identification information thus minimizing the residual risk. The HL7 agent sends the patient identifying information as their hash values along with other demographic and EHR data in an HL7 message. The steps involved in duplicate detection algorithm employed at the resource fusion layer when incoming HL7 messages with hashed identifying attributes are received are shown in Fig.8. The linking agent compares each of the patient identification value individually with the existing record values. If the attribute matches, the linked value is 1 otherwise it is 0 as shown in equation (1). The rows exceeding a threshold of number of 1s are extracted; implying that they have maximum number of common matching identifying attributes. These rows are then used in classification algorithm as matched or unmatched by a decision tree algorithm. We use Secure Hash Algorithm (SHA) that is a one-way cryptographic function and converts any data type to a fixed length value called hash value with a very low probability of two values hashing to the same value. These values are used to protect patients privacy while exchange sensitive health information. These hashed values are then utilized by the linking algorithm to fuse data at the population data bank.

$$MI_{ik} = \sum_{j=1}^{n} A_{ij} \oplus a_{kj} \quad \forall i \in \{0, .., M\} \, \forall k \in \{0, .., N\} \tag{1}$$

where j is the attribute being matched such as first name, last name, dob, age etc. and n is the total number of attributes



(a) The dynamic message interaction model shows the initiating and response ADT messages exchanged between glue layer HL7 ADT gateway and resource fusion layer HL7 ADT gateway.

```
MSH|^~\&|GLG|RFG|STHA|ShifaTest|20190415110403||ACK^A01^MDM_T01|MSG00001|P|2.7
MSA|AA|MSG00001
```

(b) A sample ACK message from ADT receiver channel to center channel.

FIGURE 7. **Transmission protocol for ADT message from source site to resource fusion layer.**

being matched. The equation shows that the *ith* row of incoming HL7 message digest is compared with *kth* row of existing population. The comparison is done attribute by attribute for matching hashed values strings for record *A* in incoming message *a* in existing databank. The value of MI_{ik} is a string of 0s and 1s with 1s representing the matched hash values and 0s representing unmatched hashed values. The rows where $MI_{ik} > threshold$ are input to the C4.5 version of decision tree algorithm that performs entropy based splitting of attributes at each node. The algorithm is then able to classify each incoming record as unique or duplicate in the fusion layer consolidated database.

5) CONSOLIDATED DATABASE

The consolidated population data bank contains several tables for EHR and other contextual data. This data is HL7 compliant and is cleaned and transformed in a unified table structure at the source glue layer. The tables are populated from the data received from the HL7 gateway. For efficient search and querying of data, this database is created in a relational manner. The tables contained in this databank are designed to populate the specialized data marts based on ETL processes or OLAP. We take a consolidated database to data mart approach instead of data marts to data warehouse approach, as later provides several benefits over the former in terms of public health analytics. For example multiple source

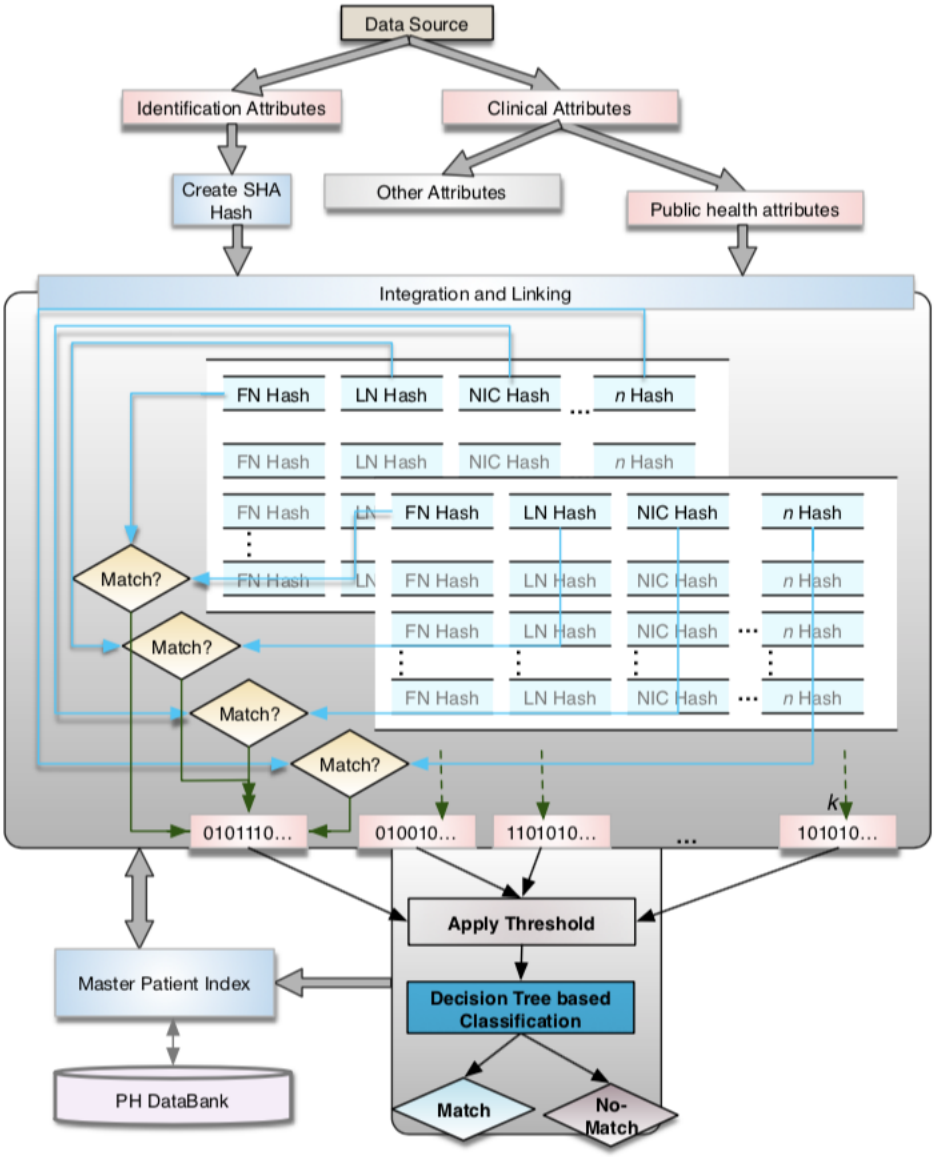


FIGURE 8. The steps involved in the algorithm for duplicate detection in incoming hashed demographic attributes and pre-existing fusion layer records. Each hashed value is compared with the corresponding attribute's existing hashed value, if the values are same, a match bit- 1 is generated otherwise 0 is set for that position. This process generates a sequence of bits which are then used as input to decision tree process.

systems are sending redundant data to populate several data marts if data marts to data warehouse approach is used where, for every spawned data mart, the incoming HL7 messages need to be directed to all the data marts interested in the arriving fields. Furthermore, in order to populate data marts with the incoming HL7 messages, the early binding of data must be modeled which in turn reduces the analytical capabilities at the service layer. Therefore, the consolidated database is created that offers raw data representation free from any business binding rules and is based only on an interoperability standard. Making this database HL7 compliant does not impose any restrictions on the data modeling since HL7 only offer interoperability specifications and does not impose any storage constraints. In addition, for a significantly comprehensive population database, atomic data representation need to be present, that is, every patient demographic and visit detail need to be recorded that may not be possible with data mart to database approach as data marts represent data at a level higher than the atomic data. Also included in this databank are tables for error logging and their subsequent reporting to source layer or authorized administrative personnel.

6) DATA MARTS

Multiple data marts are spawned out from the underlying consolidated population databank to support regional and national programs running under different public health departments. The data marts provide analytical capabilities on real-time or retrospective data. In addition to several national program data marts, evaluation data marts, data quality data marts, insurance and claims data marts can also be supported. The availability of specialized relatively smaller sized data marts ensures accurate and search efficient data availability to authorized public health departments or researchers. These data marts are created by extracting data from population bank using Extract Load Transform (ELT) process or by performing transformations using joins or Online Analytical processing (OLAP). The data marts are the eventual sources of data requested by the service layers under different programs. These marts are refreshed based on underlying data changes where real time data is required or program appropriate snapshots of data are taken where retrospective data is required. Some examples of the data marts are surveillance data marts that might eventually replace the surveillance reports, research data marts with de-identified clinical data for research purpose

In addition to these components, implementation based integrated components are included in PHF. For example, a query engine allows to express meaningful operations and exploit the powerful data stored in the public databank. At resource fusion layer a population wide data bank exist that includes information from multiple, heterogeneous and varied sources in terms of format, privacy and access level. Therefore, the query engine at this layer must be able to pull and integrate the required information from the underlying multitudes of data to create data marts or allow direct access to the population databank. An authorization server provides role based access to the data marts and the consolidated database. Anonymization agent performs the appropriate de identification and of health data when making it available for the research purpose in accordance with the HIPAA privacy rules. Context resolution agent is responsible for handling incoming data from sources other than health for example, population census data, agriculture data, economic data etc and link it with the existing data zones. The internal design of these components are not considered under the scope of this article.

V. ANALYSIS AND DISCUSSION

This section discusses the afore mentioned objectives of PHF in section IV in terms of their implementation details and their significance for public health analytics. In the light of the following discussion, it can be concluded that PHF is able to assist evidence based decision making by providing scientific evidence on the basis of available EHR and contextual resources.

A. REPRESENTATION OF PRIVATE SECTOR IN NATIONAL HEALTH

In Pakistan public health perspective, PHF allows participation of private health care facilities in public health programs. The public health programs in Pakistan are vertical and fragmented that run in parallel and include surveillance data from public sector mostly. Only a small part of private sector health care facilities offers to share health data due to privacy and data protection risks. Consequently, health data fail to represent the population visiting the private sector care providers and valuable volumes of data gets omitted during public health analysis and eventually policy decisions. While, only a small fraction of public sector health facilities use automated EHR systems due to lack of resources and funding, many private health service providers use automated and autonomous EHR systems. In addition, quality of service and consequently data present in private sector is better than available from public sector hospitals. The legacy systems are not required to fully migrate their data from current environment to HL7 compliant environment. They only need to make the data accessible that is of public health interest in an anonymized form. The responsibility of data anonymization also lies with the framework. Therefore, in the presence of appropriate regulations, this allows the public as well as private health care facilities to participate in national public health programs by providing evidence for decision making process.

B. LEGACY SYSTEMS SUPPORT

In Pakistan the legacy EHR systems are either non-existent or autonomously developed and implemented in the private sector health care facilities. They exchange their data with external systems or non technical entities through faxes, email or even paper based reports. In general, they are not designed for information sharing or health data exchange despite the many vertical reporting health programs running under the ministry of health. Additionally, the legacy systems are designed around the workflow and architectural requirements of internal functioning of a health care facility. Consequently, the changes required to be implemented in legacy EHR systems in order to support health data exchange with external entities are high cost, time consuming or technically infeasible. PHF is able to support such systems in order to move heath data of public health interest that allows the use of data of public health interest without having to fully migrate from existing systems to HL7 compliant system. through the implementation of glue layer. The glue layer acts as a middleware between legacy EHR systems and public health resource server through following process: 1) It extracts the EHR features of public health interest from the data repository at the participating health care facility. This extraction can be done based on a predefined interval, event or request from the upper layers through HL7 gateway. 2) It provides an anonymization agent that converts the patient identifying information to irreversible hashed codes.3) It implements an HL7 agent that converts the non standard EHR format data into HL7 specification based segments and stores in the local cloud. 4) It implements an HL7 gateway that send information from the local cloud to the public health resource server cloud. This allows minimum intervention of legacy systems in order to participate in the public health support data programs. Consequently, the infrastructure elements are able to work together to support legacy without and do not impose the need for building from scratch.

C. STANDARD BASED INTEROPERABILITY

The glue layer is an abstraction layer operating over an existing EHR system of health care facility and is comprised of independent interoperability components. These components are developed to create a standard semantic encoding of the health data obtained from the legacy system. The underlying multi format and heterogeneous systems can be based on relational DBMS, access or simple file management systems, the data of interest, in our case, data required for public health analytics is extracted from the system and converted into a unified format for forwarding to upper layers. For this purpose, we chose HL7 v.2 as it offers large set of messages, documents and services that can facilitate the process of creating interfaces to interact with the legacy system as well as address the public health data requirement challenges.

D. PRIVACY PRESERVATION

In public health perspective, the identity of a patient is not significant. However, in order to build a definable population health data bank, the patients record must be fused or integrated in a seamless manner. While the use of EHR for evidence based policy and decision making at public health level has been repeatedly studied and its importance emphasized, there are certain barriers and challenges in achieving this goal. PHF offers privacy preserving linking of patient records where patient demographic attributes are not fully disclosed. This is done through SHA 256 hash function and a novel linking algorithm discussed in section IV-B.4

E. INCLUSION OF CONTEXT

Many factors are associated with improved health care delivery services. Factors such as economic stability, patients' ability to access quality care, education, linguistic and cultural issues, climate and environmental context, have been studied in literature and are known as significant social determinants of public health outcomes. However, very few studies have included these context indicators in public health surveillance based frameworks. In addition, context may also include cost incurred on various pubic health programs and their respective subtasks such as data collection, analysis and implementation. The inclusion of this information is crucial for assessment of evidence about public health interventions for policy decision making. The absence of the context from the existing framework is due to their absence or unavailability at the original evidence base that in turn maybe due to the fact that these programs do not have the primary objective of identifying evidence on implementation, cost and sustainability. This means that this context is generally not available to the policy makers in combination with population health status who need to make decisions. PHF is able to make this context information available to both researchers and decision makers through the resource fusion layer.

This layer makes the context available that is already being acquired and stored by other departments under different programs in silos for their particular intent with some of the parameters being publicly available, such as weather data, population census data, urbanization data etc. Therefore, there is no ethical concern of using this already collected data for its useful inclusion in public health context and for generating annotated datasets for research purposes.

F. INCLUSION OF RESEARCH OUTCOMES FOR EVIDENCE BASED DECISION MAKING

Traditionally, for research purposes data is prospectively collected in random clinical environments which provides a control over data definitions and pre processing steps. On the contrary, the data public health surveillance data includes only the disease incidences based on the diagnosis or laboratory tests while records with negative test results are excluded from the reports. This happens because data is collected under a specific public health program and minimum required data approach is encouraged where the survey targets a specific health state, disease, or population type. However, for a meaningful pattern recognition from health datasets using AI and ML techniques, it is important to have annotated datasets and not just positive incidence reports. PHF allows the creation, anonymization and publication of multi domain labeled health datasets that can be accessed by researchers through creation of data marts. This will assist in creating predictive models for disease surveillance that can then be implemented to be used by public health authorities. The models will make it possible to include research outcomes in public health evidence based decision making.

G. SCALABILITY

The glue layer is designed in a way as to integrate heterogeneous data formats. For every new EHR data format only the mapping component has to be redesigned in order to scale horizontally at the evidence base level. More health care facilities can be added to the framework by implementing glue layer at the local end. Scalability is supported in terms of adding more clinical or EHR sources, integrating more contextual sources as well as scaling at analytical end to support more decision support systems. The system is currently being implemented as a centralized server, however, due to modular architecture of PHF, it is possible to migrate to a distributed cloud approach. In regard to data security in distributed environment, the data stored is anonymized using irreversible hash algorithm and can be scaled to a cloud based storage because of effective linking algorithm presented. In addition, security protocols are implemented on role based access of data marts to support increasing number of end users.

VI. CONCLUSION

Health care is a global problem and needs solutions that are based on standard specifications. The EHR system have emerged as a useful evidence base in public health informatics in addition to its primary purpose of electronic medical data storage. We have presented a framework based on standardized acquisition and transmission of EHR that can be applied worldwide even to a fragmented public health systems that currently work in silos with their own data formats and analysis techniques. The framework called Public Health Framework(PHF) defines a layered architecture where data is transmitted from EHR to policy making level that then become part of regulation ensuring that data related to public health should be pushed to upper layers based on appropriate triggers. The framework is valid where either the digital EHR is available or paper based EHR has to be digitalized for incorporation into national health care framework. The only limitation of the PHF is that requires adaptors at individual healthcare facilities for extracting data relating to public health, but such adaptors can easily be developed and deployed to ensure regulatory compliance by governments in ensuring all public health data sources, whether private or public, are represented and public health is given due importance in policy making. In future PHF will be offered as an open source resource for adoption by other countries or state health care entities. PHF is intended for improving population health by disease prevention and outbreak detection focused analysis. Implementing this framework at local, state and national level will help connect patients, healthcare practitioners and public health authorities to conduct a joint effort to reshape public health surveillance and its outcomes.

ACKNOWLEDGMENT

The authors thank Shifa International Hospital, Islamabad, Pakistan for providing a platform for validation of framework on real patient data. This work is part of PM Task Force on IT and Telecom initiative for population health care improvement and is presented as an initial proposed framework to the taskforce.

REFERENCES

[1] G. B. Laleci, M. Yuksel, and A. Dogac, "Providing semantic interoperability between clinical care and clinical research domains," *IEEE J. Biomed. Health Informat.*, vol. 17, no. 2, pp. 356–369, Mar. 2013.

[2] G. S. Birkhead, M. Klompas, and N. R. Shah, "Uses of electronic health records for public health surveillance to advance public health," *Annu. Rev. Public Health*, vol. 36, pp. 345–359, Mar. 2015.

[3] B. Shickel, P. J. Tighe, A. Bihorac, and P. Rashidi, "Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis," *IEEE J. Biomed. Health Inform.*, vol. 22, no. 5, pp. 1589–1604, Sep. 2018.

[4] O. Jacobson and H. Dalianis, "Applying deep learning on electronic health records in swedish to predict healthcare-associated infections," in *Proc. BioNLP@ACL*, 2016, pp. 191–195.

[5] D. B. Neill, "New directions in artificial intelligence for public health surveillance," *IEEE Intell. Syst.*, vol. 27, no. 1, pp. 56–59, Jan./Feb. 2012.

[6] X. Luo and S. Li, "Non-negativity constrained missing data estimation for high-dimensional and sparse matrices," in *Proc. 13th IEEE Conf. Automat. Sci. Eng. (CASE)*, Aug. 2017, pp. 1368–1373.

[7] L. Hu, X. Yuan, X. Liu, S. Xiong, and X. Luo, "Efficiently detecting protein complexes from protein interaction networks via alternating direction method of multipliers," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, to be published.

[8] Q. A. Qureshi, N. A Qureshi, M. Z. Khan, A. Nawaz, and B. Shah, "Issues and prospects of e-health in Pakistan," *Medit. J. Med. Sci.*, vol. 1, pp. 31–52, Oct. 2014.

[9] A. F. Klaib and M. S. Nuser, "Evaluating EHR and health care in Jordan according to the international health metrics network (HMN) framework and standards: A case study of Hakeem," *IEEE Access*, vol. 7, pp. 51457–51465, 2019.

[10] D. G. Katehakis, A. Kouroubali, and I. Fundulaki, "Towards the development of a national ehealth interoperability framework to address public health challenges in greece," in *Proc. SWH@ISWC*, 2018, pp. 1–9.

[11] M. S. Qazi and M. Ali, "Pakistan's health management information system: Health managers' perspectives," *J. Pakistan Med. Assoc.*, vol. 59, no. 1, pp. 10–14, Jan. 2009.

[12] R. Kumar, B. T. Shaikh, A. K. Chandio, and J. Ahmed, "Role of health management information system (HMIS) in disease reporting in a rural district of Sindh," *Pakistan J. Public Health*, vol. 2, no. 2, pp. 10–12, Jun. 2012.

[13] I. M. Baytas, K. Lin, F. Wang, A. K. Jain, and J. Zhou, "Phenotree: Interactive visual analytics for hierarchical phenotyping from large-scale electronic health records," *IEEE Trans. Multimedia*, vol. 18, no. 11, pp. 2257–2270, Nov. 2016.

[14] S. E. Perlman, K. H. McVeigh, L. E. Thorpe, L. Jacobson, C. M. Greene, and R. C. Gwynn, "Innovations in population health surveillance: Using electronic health records for chronic disease surveillance," *Amer. J. Public Health*, vol. 107, no. 6, pp. 853–857, 2017. doi: 10.2105/AJPH.2017.303813.

[15] M. Bates, "Tracking disease: Digital epidemiology offers new promise in predicting outbreaks," *IEEE Pulse*, vol. 8, no. 1, pp. 18–22, Jan./Feb. 2017.

[16] Health Leven Seven International. *Introduction to HL7 Standards*. Accessed: Jun. 23, 2017. [Online]. Available: http://www.hl7.org/implement/standards/index.cfm?ref=nav

[17] J. Lombardo, H. Burkom, E. Elbert, S. Magruder, S. H. Lewis, W. Loschen, J. Sari, C. Sniegoski, R. Wojcik, and J. Pavlin, "A systems overview of the electronic surveillance system for the early notification of community-based epidemics (ESSENCE II)," *J. Urban Health*, vol. 80, no. 1, pp. i32–i42, Mar. 2003. doi: 10.1007/PL00022313.

[18] P. J.-H. Hu, D. Zeng, H. Chen, C. Larson, W. Chang, C. Tseng, and J. Ma, "System for infectious disease information sharing and analysis: Design and evaluation," *IEEE Trans. Inf. Technol. Biomed.*, vol. 11, no. 4, pp. 483–492, Jul. 2007.

[19] Y. Zhang, Y. Dang, Y.-D. Chen, H.-C. Chen, M. Thurmond, C.-C. King, D. D. Zeng, and C. Larson, "BioPortal infectious disease informatics research: Disease surveillance and situational awareness," in *Proc. Int. Conf. Digit. Government Res.*, May 2008, pp. 393–394.

[20] F.-C. Tsui, J. U. Espino, V. M. Dato, P. H. Gesteland, J. Hutman, and M. M. Wagner, "Technical description of RODS: A real-time public health surveillance system," *J. Amer. Med. Inform. Assoc.*, vol. 10, no. 5, pp. 399–408, 2003.

[21] F.-C. Tsui, J. U. Espino, M. M. Wagner, P. Gesteland, O. Ivanov, R. T. Olszewski, Z. Liu, X. Zeng, W. Chapman, W. K. Wong, and A. Moore, "Data, network, and application: Technical description of the UTAH rods winter olympic biosurveillance system," in *Proc. AMIA Symp.*, Feb. 2002, pp. 815–819.

[22] D. W. Gould, D. Walker, and P. W. Yoon, "The evolution of BioSense: Lessons learned and future directions," *Public Health Rep.*, vol. 132, no. 1, pp. 7S–11S, 2017. doi: 10.1177/0033354917706954.

[23] M. Ginsberg, J. Johnson, J. Tokars, C. Martin, R. English, G. Rainisch, W. Lei, P. Hicks, J. Burkholder, M. Miller, K. Crosby, K. Akaka, A. Stock, and D. Sugerman, "Monitoring health effects of wildfires using the biosense system—San Diego County, California, October 2007," *Morbidity Mortality Weekly Rep.*, vol. 57, pp. 741–744, Jul. 2008.

[24] J. M. Overhage, S. Grannis, and C. J. McDonald, "A comparison of the completeness and timeliness of automated electronic laboratory reporting and spontaneous reporting of notifiable conditions," *Amer. J. Public Health*, vol. 98, no. 2, pp. 344–350, Feb. 2008.

[25] S. Bowman, "Impact of electronic health record systems on information integrity: Quality and safety implications," *Perspect. Health Inf. Manage.*, vol. 10, p. 1c, Oct. 2013.

[26] M. Klompas, N. M. Cocoros, J. T. Menchaca, D. Erani, E. Hafer, B. Herrick, M. Josephson, M. Lee, M. D. P. Weiss, B. Zambarano, K. R. Eberhardt, J. Malenfant, L. Nasuti, and T. Land, "State and local chronic disease surveillance using electronic health record systems," *Amer. J. Public Health*, vol. 107, pp. 1406–1412, Sep. 2017.

[27] M. Klompas, J. McVetta, R. Lazarus, E. Eggleston, G. Haney, B. A. Kruskal, W. K. Yih, P. Daly, P. Oppedisano, B. Beagan, M. Lee, C. Kirby, D. Heisey-Grove, A. DeMaria, and R. Platt, "Integrating clinical practice and public health surveillance using electronic medical record systems," *Amer. J. Preventive Med.*, vol. 42, pp. S154–S162, Jun. 2012.

[28] E. Samoff, M. T. Fangman, A. T. Fleischauer, A. E. Waller, and P. D. Macdonald, "Improvements in timeliness resulting from implementation of electronic laboratory reporting and an electronic disease surveillance system," *Public Health Rep.*, vol. 128, no. 5, pp. 393–398, 2013.

[29] N. G. Weiskopf, G. Hripcsak, S. Swaminathan, and C. Weng, "Defining and measuring completeness of electronic health records for secondary use," *J. Biomed. Inform.*, vol. 46, no. 5, pp. 830–836, Oct. 2013.

[30] K. Baltrusaitis, J. S. Brownstein, S. V. Scarpino, E. Bakota, A. W. Crawley, G. Conidi, J. Gunn, J. Gray, A. Zink, and M. Santillana, "Comparison of crowd-sourced, electronic health records based, and traditional healthcare based influenza-tracking systems at multiple spatial resolutions in the United States of America," *BMC Infectious Diseases*, vol. 18, no. 1, p. 403, Aug. 2018. doi: 10.1186/s12879-018-3322-3.

[31] M. Santillana, A. T. Nguyen, T. Louie, A. Zink, J. Gray, I. Sung, and J. S. Brownstein, "Cloud-based electronic health records for real-time, region-specific influenza surveillance," *Sci. Rep.*, vol. 6, May 2016, Art. no. 25732. doi: 10.1038/srep25732.

[32] P. M. Coloma, G. Trifirò, M. J. Schuemie, R. Gini, R. Herings, J. Hippisley-Cox, G. Mazzaglia, G. Picelli, G. Corrao, L. Pedersen, J. van der Lei, and M. Sturkenboom, "Electronic healthcare databases for active drug safety surveillance: Is there enough leverage?" *Pharmacoepidemiol. Drug Saf.*, vol. 21, pp. 611–621, Jun. 2012.

[33] M. S. Calderwood, R. Platt, X. Hou, J. Malenfant, G. Haney, B. Kruskal, R. Lazarus, and M. Klompas, "Real-time surveillance for tuberculosis using electronic health record data from an ambulatory practice in eastern Massachusetts," *Public Health Rep.*, vol. 125, no. 6, pp. 843–850, 2010.

[34] J. Zhao, P. Papapetrou, L. Asker, and H. Boström, "Learning from heterogeneous temporal data in electronic health records," *J. Biomed. Inform.*, vol. 65, pp. 105–119, Jan. 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1532046416301654

[35] H. Chen, D. Zeng, and P. Yan, *Infectious Disease Informatics: Syndromic Surveillance for Public Health and Bio-Defense* (Integrated Series in Information Systems), vol. 67. Springer, 2009, p. 2.

[36] G. J. Milinovich, S. M. R. Avril, A. C. A. Clements, J. S. Brownstein, S. Tong, and W. Hu, "Using Internet search queries for infectious disease surveillance: Screening diseases for suitability," *BMC Infectious Diseases*, vol. 14, p. 690, Dec. 2014.

[37] R. Kukafka, J. S. Ancker, C. Chan, J. Chelico, S. Khan, S. Mortoti, K. Natarajan, K. Presley, and K. Stephens, "Redesigning electronic health record systems to support public health," *J. Biomed. Inform.*, vol. 40, no. 4, pp. 398–409, 2007. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1532046407000603

[38] A. Tomines, H. Readhead, A. Readhead, and S. Teutsch, "Applications of electronic health information in public health: Uses, opportunities & barriers," *eGEMs*, vol. 1, no. 2, p. 1019, 2013.

[39] A. F. Elliott, A. Davidson, F. Lum, M. F. Chiang, J. B. Saaddine, X. Zhang, J. E. Crews, and C.-F. Chou, "Use of electronic health records and administrative data for public health surveillance of eye health and vision-related conditions in the United States," *Amer. J. Ophthalmol.*, vol. 154, no. 6, pp. S63–S70, Dec. 2012.

[40] Y. Yang, X. Li, N. Qamar, P. Liu, W. Ke, B. Shen, and Z. Liu, "Medshare: A novel hybrid cloud for medical resource sharing among autonomous healthcare providers," *IEEE Access*, vol. 6, pp. 46949–46961, 2018.

[41] E. C. Schiza, T. C. Kyprianou, N. Petkov, and C. N. Schizas, "Proposal for an eHealth based ecosystem serving national healthcare," *IEEE J. Biomed. Health Inform.*, vol. 23, no. 3, pp. 1346–1357, May 2019.

[42] E. Rahim, "An overview of the health an overview of health sector: The way forward," Multi-donor Support Unit, Ministry Health, Government Pakistan, Tech. Rep. 30611, Nov. 2001. [Online]. Available: http://documents.worldbank.org/curated/en/819181468758747563/pdf/30611.pdf

[43] E. Rahim, "An overview of the health an overview of health sector: The way forward," Ministry Health, Government Pakistan, Multi-Donor Support Unit, Islamabad, Pakistan, Tech. Rep. 30611, Nov. 2001. [Online]. Available: http://documents.worldbank.org/curated/en/819181468758747563/pdf/30611.pdf

[44] JIC. Agency. (Apr. 2009). *The District Health Information System (DHIS) Project for Evidence-Based Decision Making and Management*. [Online]. Available: https://www.jica.go.jp/english/index.html

[45] (Jan. 2019). *WHO Presence in Pakistan*. [Online]. Available: http://www.emro.who.int/pak/programmes/health-managment-information-system.html

[46] *National Health Accounts for Pakistan*, Ministry Health, Government Pakistan, Pakistan Bureau Statist., Islamabad, Pakistan, 2016.

[47] Ministry of National Health Services, Regulations and Coordination, Government Pakistan. (Aug. 2016). *National Health Vision, Pakistan*. [Online]. Available: http://www.nhsrc.gov.pk/index.php?page=publicinfo

[48] HLS International. *Hl7 Version 2 Product Suite. Health Level Seven International*. [Online]. Available: https://www.hl7.org/implement/standards/product_brief.cfm?product_id=185

[49] F. Khalique, S. A. Khan, Q.-U.-A. Mubarak, and H. Safdar, "Decision tree-based anonymized electronic health record fusion for public health informatics," in *Proc. Sci. Inf. Conf.*, 2018, pp. 404–414.

FATIMA KHALIQUE received the M.S. degree in computer science from Uppsala University, Sweden, in 2007. She is an Oracle Certified Professional and has worked in industry as well as academia. She was a Lecturer with the National University of Sciences and Technology (NUST) and the National University of Modern Languages (NUML), Islamabad, Pakistan. She was also a Software Developer with Zhonxing Telecom Engineering (ZTE), Islamabad. She is currently a Ph.D. Scholar with the National University of Sciences and Technology (NUST), Islamabad. Her research interests include data mining, health informatics, computer networks, system design and testing, and machine learning algorithms.

SHOAB A. KHAN received the Ph.D. degree in electrical and computer engineering from the Georgia Institute of Technology, Atlanta, GA, USA. He is currently a Professor of computer and software engineering with the College of Electrical and Mechanical Engineering, National University of Sciences and Technology (NUST). He is an inventor of five awarded U.S. patents and has over 260 international publications. His book on digital design is published by John Wiley & Sons and is being followed in national and international universities. He has more than 22 years of industrial experience in companies in USA and Pakistan. He received Tamgh-e-Imtiaz (Civil), the Highest National Civil Award in Pakistan, the National Education Award, in 2001, and the NCR National Excellence Award in Engineering Education. He is the Founder of the Center for Advanced Studies in Engineering (CASE) and the Center for Advanced Research in Engineering (CARE). CASE is a primer engineering institution that runs one of the largest postgraduate engineering programs in the country and has already graduated 50 Ph.D. students and more than 1800 M.S. students in different disciplines in engineering, whereas CARE, under his leadership, has risen to be one of the most profound high technology engineering organizations in Pakistan developing critical technologies worth millions of dollars for organizations in Pakistan. CARE has made history by winning 13 PASHA ICT awards and 11 Asia Pacific ICT Alliance Silver and Gold Merit Awards while competing with the best products from advanced countries like Australia, Singapore, Hong Kong, Malaysia, and so on. He has served as the Chairman of the Pakistan Association of Software Houses (P@SHA) and as a member of the Board of Governance of many entities in the Ministry of IT and Commerce. He has also served as a member of the National Computing Council and the National Curriculum Review Committee.

IRUM NOSHEEN received the B.S. degree in software engineering from Fatima Jinnah Women University, Rawalpindi, Pakistan, and the M.S. degree in computer engineering from the University of Engineering and Technology, Taxila, Pakistan, in 2009. She has been a Research Assistant with the Center for Advanced Research in Engineering (CARE), since 2012. She has also been a Lecturer with the Department of Electrical Engineering, Faculty of Engineering and Technology, International Islamic University, Islamabad, Pakistan, since 2009. She is currently a Ph.D. Scholar with the Center for Advanced Studies in Engineering (CASE). Her research interests include computer networks, infrastructure-less networks, communication in mission, and time critical networks.