DOI: http://doi.org/10.4038/icter.v13i3.7221

International Journal on Advances in ICT for Emerging Regions 2020 13(3):

A Multi-layered Data Preparation Model for Health Information in Sudan

Ahmed Mustafa Abd-Alrhman[#1], Love Ekenberg [#2]

***Abstract*— Data quality is a major challenge in almost every data project in today's world, especially when the required data has a national or global look and feel. However, data preparation activities dominate the efforts, cost, and time consumption. Nowadays, many data collection approaches are continuing to evolve in the era of big data to accommodate revolutionary data flows, especially in the health sector, which contains many different levels of data types, formats, and structures. The lack of qualified and reliable data models is still an ongoing challenge. These issues are even magnified in developing countries where there is a struggle to make advances in health systems with limited resources environments, and to adopt the advantages of ICT to minimize the gaps in health information systems. This article introduces a geo-political multi-layered model for data collection and preparation, the model enables the health data to be collected, prepared and aggregated by using data attendance approach and address data challenges such as data missing, incompetence and format. The currently used data collection method in health sector in Sudan was analysed and data challenges were identified, with respect to geo-political structure of the country. The result of the model provides structured datasets framed by time and geographical spaces that can be used to enrich analytical projects and decision-making in the health sector.**

***Keywords*—Data preparation, Data quality, Sudan, Health Information systems**

I. Introduction

In today's world, the use of data is a common necessity in almost every domain, including individuals, groups, enterprises, and governmental entities; however, different usages and needs characterize the data type, format, and volume. Data management introduces an entire span of sequential processes from the generation of a particular dataset until archiving or in some cases deletion; this includes generation, collection, recording, processing, transforming, and loading of data, known as Data Lifecycle Management (DLM) [1]. Many developing countries, including Sudan, suffer from data collection problems, particularly in the health sector [2-4]. In the case of Sudan, the governance body for health, which is the Federal Ministry of Health (FMoH), obligates health providers (private, public) to record and submit a monthly manual health report which summarizes different medical and management statistical data in any health facility. The report is structured into subsections in a unified fashion for all health service providers. Each section contains a list of heading titles and blank spaces for required data entries, for the purpose to record the existence of the services in the facility, in addition to numerical values that represent: services, tests, patients, equipment, and personnel. The analysis of the data herein pointed out many challenges, for instance, some reports have 35% incomplete data, other reports had more than 25 manual corrections, in addition to 5-7% of missing data and inconsistencies between working electronic laboratory systems and the corresponding manual aggregated data in the monthly reports for one health facility. With this reservation, the results generally showed many challenges in the data collection and management, which are categorized into three problematic themes: data, user, and management errors.

- The main data errors are (1) missing data: some columns are left entirely blank, without a trace for default values; (2) incompleteness: some patient files were found not fully recorded; (3) inconsistent data: for example, this was found when comparing existing software results with designated data elements in the report and variances were identified; and (4) unavailability: in some cases, the data was not ready in data collection points at reporting time.
- The interaction of the user (data reporter) with the medical report in terms of collection, recording, and submission of data indicates issues such as (1) manual data correction or cancellation using stationary tools; (2) collection of data being done verbally and not documented in some cases; (3) the collected data were not supervised in some cases, as the collection department consists of one person who is responsible for all reporting activities; and (4) harmonization problems between medical and management staff in data collection.
- The management of data suffers from many obstacles including: (1) manual collection of data, which might be exposed to human mistakes; (2) the absence of validation or verification mechanisms; (3) poor time constraints; for instance, data may be reported as monthly aggregations, and reports can be delayed for 10 days from submission dates; (4) poor data structure, with no standard coding mechanisms and no data links, which overburdens future analysis.

The overall results indicate poor data quality, which is considered one of biggest challenges facing those who working with data preparation and management. Despite the notion that the time spent in collecting and preparing data to ensure its quality dominates total efforts and represents almost 60% of the duration of data analysis projects [5], applying ICT in health domain can have a positive impact on healthcare provision in developing countries, especially with the increases in network coverage and data processing

Manuscript received on 22 Sep. 2019. Recommended by Dr. Manjusri Wickramasinghe on 20 November 2020.

Ahmed Mustafa Abd-Alrhman,is from Sudan University For Science and Technology. (*ahmedabudi@hotmail.com*).

Love Ekenberg is from the Department of Computer and Systems Sciences, Stockholm University (*lovek@dsv.su.se*) and the International Institute for Applied Systems Analysis (IIASA) (*ekenberg@iiasa.ac.at*)

capacities and the decreases in hardware costs. However, challenges remain outstanding to develop data models that can satisfy data quality measures and management, to support and strengthen health information systems.

Data usually needs further processing in order to be ready for use, that is, to transform the collected data into other shapes and formats to meet new requirements such as data analysis. This has become even more necessary in the case of non-electronic data forms, which put and additional preliminary step to transform the data into a soft version as an essential requirement to enhance the upcoming data management [6], as noted by Sarkies et al. [7] who estimated that an improvement of more than 30% could be achieved by using an electronic patient recording program rather than manually recording patient data.

In this article, we suggest an innovative approach to build structured data collection systems for health information in Sudan and elsewhere whit similar conditions. We are suggesting a geo-political data hosting approach to systematically organize the data according to demographical dimensions with political control mechanisms. The reason for choosing such a structure is to better support the hierarchy of official health authorities, which are responsible for maintaining policies, regulations, and management activities. We introduce a data model to structure data sources using geo-political settings to represent health organizations in Sudan and provide a mechanism for exchanging transactional and aggregated data, between health facilities and governing health bodies. We introduce a different approach in the data preparation process by moving the processes of data preparation activities downstream to data collection points, compared to traditional approaches that process and qualify the data after completion of data collection phase. The objectives of the proposed approach mainly focus on:

- Integrating data preparation techniques in early stages at the data entry point to increase the efficiency and speed of data preparation process;
- Early detection of errors and isolation of low data quality and fix data errors at data entry points;
- The new approach transfers part of the process of Extract, Transform, Load (ETL) to data entry points in order to enhance the load balance, process throughput, and data production time.

The next section will discuss data management in the health sector, and highlight the data quality requirements and its impact on health information, after, we will describe the current status of ICT implementations in Sudan and identify the requirements from health data models in this developing country. The third section introduces the Multi-layered Data Attendance Collection Model (MDACM) and its configuration. Lastly, a discussion is presented to show the model's ability to function in an environment with a poor ICT infrastructure.

II. Methodology

The intention was to understand the data collection problems inside the health facilities as well as the communication mechanisms and challenges with health authorities and we have primarily used a qualitative approach. An analysis of the monthly manual reports was conducted to investigate the report structure, contents, and user interaction possibilities. The analysis included five large hospitals in the capital city, Khartoum, whereof two were public and three in private ownership. The selection of health facilities was constrained in terms of the service capacity and frequency of patients. Monthly reports were analysed for three months (June - August 2018) for each hospital. To that, digital database analysis was conducted in the health facilities using digitalised systems to verify the manual monthly report data and to identify any possible discrepancies. Furthermore, interviews were conducted with management employees, responsible for generating the monthly reports. Moreover, interviews with medical staff were conducted to identify the possible gaps and obstacles in the data collection processes. We also monitored the actual data collection, verification and exchange to validate this information.

III. Literature Review

A. Data Management Challenges in Health Sector

It is obvious that data management in the health sector is crucial because of the sensitivity of the domain and its tight relationship with society, economics, security, and public health. However, the importance of data preparation extends the perception of data usage to decreasing operational cost for organizations and better resource management. A cost-benefit analysis of data quality conducted by Redman [8] estimated that data errors in the range of 1–5% can cause revenues losses by 10% in industry. This can also be a burden in developed countries; for instance, Eckerson [9] estimates that the cost to US businesses of poor data quality estimated by 600 billion USD per year. It is not only the financial cost that is negatively impacted by poor data quality, data analytics and the IT industry also share such concerns; for instance, a survey conducted by a professional data preparation company [10] showed that data preparation costs almost 450 billion USD per year. Even in such circumstances, the survey indicates that data analysts prefer to work with modelling and analytic activities rather than spending time preparing data.

The impact of poor data quality can affect health provision, as noted by Yawson and Ellingsen who investigated the implementation of an electronic health record system in Ghana, where they concluded that recording huge amounts of data in health information systems does not necessarily indicate an improvement in health provision quality because of the bad quality of the data [11]. This reveals that data models should be strengthened with quality measures in the data collection and acquisition phases to produce a valid and qualified dataset. Collected data might contain errors and the quality of data can be negatively impacted by, for example, incompleteness, inconsistency, and duplications, which can be a challenge for data projects in various phases. Rahm and Hai Do noticed the problem of data quality when uploading data into data warehouses that can contain highly probable quality issues occurred at their original sources [12]. This is shown in the process of data cleansing, which occurs after data collection. The problem observed here, in the concentration and delays of the work burden until the collection process is finished. In the same context, Zhang highlight the importance of the data quality in data preparation [13], which can speed up performance of the data analysis and mining processes. The authors identify the needs to purify and clean the collected data, which might be collected from different sources with multiple issues such as incompleteness and inconsistency [13]. Moreover, Kwak and Kim pointed the problems of contextual missing values and

the significance of outliers' affection in statistics estimation [14]. Aguinis et.al discussed and recommended the best data preparation techniques and addressed data issues, such as outliers management, data correction methods and data transformation [15].

The density of data in health is always characterized by the accumulation of a huge amount of data which is batched as datasets. Many researchers recognize probable obstacles in managing big datasets and many solutions and approaches have been proposed in that context. For instance, Kanchi et al. identify the challenges when managing big data including health records and recommended optimization methods to reduce data problems by concentrating on data management practices, techniques, and infrastructure, which can have a positive impact on data quality [16]. In the same context, Levant et al. suggest that data quality of systems can be improved significantly by addressing the data problems and the consequent remedies, in particular, data correctness, data completeness, precision, timeliness, and usability [17]. From that perspective, they suggest the enhancement of systems' data quality by building semantically rich data models, increasing the database rules and constraints, and using a predefined process for data usability. The size of data, complexity and data sources can be viewed as a significant challenge as was pointed by Kristian [18].

B. Data Quality Management

Different approaches have been introduced to measure data quality. For instance, Heinrich et al. introduced a metric-based approach to quantify the data quality by adding data correctness and a timeline to meet the system requirements [19]. Rather than using pre-defined requirements as recommended by Heinrich [19], Cappiello et.al introduced another approach which uses scoring processes to compare the result outputs to match the evaluation of pre-specified objects [20]. In addition, an assessment of data quality was introduced in 2018 using the Data Quality Framework (DQF) by creating quality properties and provenance of data with respect to user's experience of the quality, to make it possible to assess the data quality and track possible data errors [21]. Svetlozar introduced some new approaches to combine data preparation with data visualization and clustering techniques to support decision-making processes [22]. Furthermore, it has been highlighted that feedbacks on collected data can prevent, e.g., incomplete data and consequently can be used for mapping and completion processing [23].

Management of health data at national levels is gaining more global attention, especially in developing countries, by bringing forward compliance with international goals [24]. Many countries including low- and middle-income countries have adopted national approaches for health data collection. For instance, in 2010, India started identifying the population eligible for state healthcare for by using Aadhaar cards with biometric verification for citizens. Similarly, Côte d'Ivoire introduced a data-sharing mobile application to monitor epidemics nationwide in 2013 [25]. The international community encourage countries for utilization of data models to improve the countries health information systems, and thus health status, which can positively enable decision makers to achieve better health provision and proper resource management, especially in severe conditions like catastrophic events and disease outbreaks.

There is global consensus on migrating from paper-based information systems to electronic format. This is especially emphasized by WHO, which describes health information systems as "the health information system provides the underpinnings for decision-making and has four key functions: data generation, compilation, analysis and synthesis, and communication and use" [26].

ICT adoption models for health information systems have been presented as systematic guidance to implement information systems based on electronic medical records. For instance, the Capability Maturity Model (CMM) is used to measure the current status of ICT implementation in health institutions by focusing on initiating and standardizing the processes with continuous improvements [27], however, the adoption of the CMM in Australia can be viewed as an enhanced version which scaled into five steps to assess the increasing capability of the e-health system in Australia. Enterprise Architecture (EA), introduced earlier by Spewak and Hill, follows a top-down approach for instantiating and improving process cycles, starting by defining enterprises requirements frameworks and governances [27]. In addition, the EMR adoption model focuses on the adoption of paperless EMR in health facilities by using seven progressive stages including a clinical decision support system, which was adopted in US and Canadian hospitals with different degrees of implementation [27]. An observation on such adoption models is that, they require a reasonable degree of stability of the ICT infrastructure, in particular, processing and storage capacities, network coverage, and personnel skills, which are the major obstacles faced by developing countries compared to developed countries. In the same context, a study by the Gates Foundation in 2009 involving 19 developing countries showed a poor ICT status, which requires these countries to adopt an alternative approach that considers the difficulties and obstacles facing ICT infrastructure.

Cost of implemention and infrastructure requirements had been identified as a real challenge in developing counties to adopt ICT in health information systems [27]. On one hand, we find the existence of ICT infrastructure in these countries; however, it is still poor and not widespread, especially network coverage; on the other hand, almost all developing countries have reasonable paper-based health information systems, which vary in terms of strength and effectiveness. Another obstacle is that without the support of reliable health information system, it is difficult to perform health management and planning nationwide [27].

Many implementations of health information systems focus on electronic patient records, although a significant improvement in this methodology can be achieved [28], however, health management, planning, and policy making; depend on disease aggregation datasets that focus on better support for medical researches and patient treatment, instead of individual patient records.

One of the biggest challenges in building electronic health information systems in developing countries is the cost of implementation [8], which requires moving paper-based data collection systems into electronic format to enable aggregation and analysis of data for the purpose of decision making, health planning, and resource management. To tackle cost and technological obstacles in developing countries, Bram et al. suggest that health data should be collected by hiring community health workers who interact

directly with the population and deliver health care and consultations [28]. The approach is constructed based on the relationship of trust between volunteers and their respective communities; however, the approach is heavily dependent on a non-stable concept, the human factor, which can be unreliable and subjective; moreover, the coverage and data management could become issues in terms of incompleteness, inconsistency, accuracy, and traceability.

Data management in developing countries, including Sudan, is therefore facing a common challenge including poor infrastructure, cost of implementation, lack of data models, and consequently poor health information systems and data quality for use it in health planning, management, and decision making.

C. Current Health Information System in Sudan

Sudan is federated as 18 states, with each state being divided into a number of localities. The population is around 40 million, distributed over 17,765,048 Km^2. Four mobile operators are licensed to operate in Sudan and they cover all the big cities and almost 83% of inhabited areas with 28 million subscribers and provide internet connectivity including 3G and 4G speeds.

Healthcare delivery in Sudan is divided between the public (governmental) and private sectors. The health management structure is divided into four organization levels. At the top is the Federal Ministry of Health (FMoH); the second level is the health ministries in the 18 states, which control – at the third level – the medical units in localities (189 localities), and health facilities are regulated at the fourth level by localities (almost 6100). There is no central record of registration of health facilities; however, states ministries are responsible for registering and licensing each health facility within their boundaries, in addition, there are referential public hospitals in the capitals of states, with extra presence in the country's capital, Khartoum.

The health information system in Sudan is an old system dated back to the twentieth century. The data has been collected from health facilities manually until the current date (June 2019). This is done by a unified monthly report for all health facilities including the private sector. The monthly form is organized as aggregated data that summarize statistics about clinical information on diseases and patient frequencies in a given facility. Each locality is responsible for aggregating and reporting, monthly, the collected medical forms under its authority to the state health ministries. Likewise, the state ministries summarize localities' aggregated data, and send it to federal ministry of health as state report in monthly or quarterly frequencies.

Data collection in the Sudanese health sector faces many challenges which negatively affect its quality. This includes many incidents of data incompleteness, inconsistency, inaccuracy, and delays in reporting time, although a prescheduled reporting times was found. In addition, the health system struggles from fragmentation and lack of a data integration mechanism between different administrations even in the FMoH. Data management is still facing major challenges, especially the process of data repairs and follow-up.

The use of ICT in the health system in Sudan is still an ongoing struggle and it has a poor status. For instance, FMoH indeed has an IT department equipped with servers, storage, and a local network with fibre optic connectivity used to connect the government's central data centre as part of an ongoing e-gov project. However, moving down the health organizational hierarchy, states and localities are poorly equipped with computers and networks, especially the medical units that are far from the central capital, with additional challenges in network coverage and power supply.

Many attempts to implement electronic health information systems have failed. However, the FMoH started to use District Health Information Software (DHIS) in 2016 for the purpose of aggregating health data from the level of localities. Implementation of the electronic system did not cover all areas, including the capital state, Khartoum, which did not adopt DHIS system because of an ongoing separate Enterprise Resource Planning (ERP) system for Khartoum state that includes a dedicated health information system for the state. In addition, other localities face technical and managerial obstacles, for example network coverage and power supply; however, the health units in the other localities vary in the level of system implementation too, and they are still reporting their monthly data twice, manually and using DHIS. IT department in FMoh prepares an annual report that reflects country's health status uses special electronic system. The report is sent to health officials, Ministry of finance, donors and international organizations. The data contained in the annual report is collected from DHIS and completed by the data collected from other health units that does not uses this system.

Sudan started to deploy a national information system to register new born in hospital, however, the system is very poor in terms of implementation and coverage. The system was implemented in private and public hospitals, but despite the presence of the system, users are still issued with birth certificates manually and some technical issues are continuing to emerge including network disconnections.

Health administration performs regular programmes of staff training, including training on the current ongoing DHIS system; however, the instability of staff is still a challenge. On the other hand, management and medical students take courses in using computers and basic operation software in universities as part of their studies. Although health workers have the basic knowledge to operate computers and basic software, capacity building is another challenge that needs more attention.

Implementation of health information systems and electronic patient records in health facilities is poor and limited. For instance, one of the three biggest public hospitals, Omdurman public teaching hospital in the capital, only uses the electronic health information system in daily work, while the other two hospitals have faced a failure to implement e-health solutions.

Private sector succeeded to use electronic systems for financial, administration and statistics purposes, especially in capital and big cities, however, there is limited attention to use applications for medical purposes. No national patient record system has been identified.

The overall status of ICT in the health sector in Sudan, as a developing country, is reasonable for it to start the adoption of a data collection model to strengthen health information and achieve continuous improvement; however, there is much that can be done to improve the use and adoption of ICT in the health sector in Sudan, especially in capacity building and IT infrastructure. To improve the quality of data and minimize data errors in the health sector in Sudan, the

data must be prepared and managed in order to be ready for analytical activities and decision making.

IV. Multi-layered Data Attendance Collection Model (MDACM)

A. Design Motivation

The data organization is intended to represent the geo-political organization of the country. The background here is that Sudan geographically is divided into 18 states, where each state has a local government, consisting of a number of local ministries. Each state is further subdivided into a number of departments, each characterized by demographic properties, where services, people, and offices are located and systematised in pre-specified geographical areas governed by the locality authorities. Therefore MDACM organizes and visualizes data in multiple layers, where each layer represents a governance level (see Figure 1). The proposed model is mainly concerned with medical data, and for that reason the model will focus mainly on providing a hierarchal data model for health organizations in Sudan.

The proposed model is intended to minimize health system obstacles by using a flexible data structure design that benefits from the current ICT infrastructure, that is, existing networks and processing and storage capacities. It is also suggesting an enabling mechanism to allow electronic data entry wherever possible while at the same time accepting data transformation from paper-based records and appropriately integrating it into the data model.

Building a national health information system should reflect the country's current health status. Health information systems in countries require data to be structured into two dimensions: time and geographical space. While a timestamp identifies the occurrence of the data, the geographical area articulates the incident locations; therefore, a time-space data representation can be provided for decision makers, health planners, and health administrators in that context.

In addition, the design of the data model is motivated by many requirements and challenges, such as support decision making, interpretation of data, and identification of patterns. Its primary focus is to collect and organize medical information to reflect the health status in the country; however, this implicitly requires modelling data to match the geo-political structure of the country, in particular, the four governing layers that are used in Sudan and represent authorities' hierarchy. By considering this outer design approach, it will provide flexibility to organize and manage health information and data flows.

Another aspect considered by the model is the concept of combining different objectives and criteria, in particular: (1) health information criteria which focus on diseases and patient medical records; (2) data management criteria, which focus on: data preparation, data quality and decision making; and (3) health management, which concerned about health provision, control health status, and better resource management. On the other hand, health information is distributed in many sources with commonalities and disparities; thus the model should organize the data to reflect the status of different health medical sections, data types, formats, and health programmes.

Stakeholders in health domain- include governmental bodies, health workers, health providers, international organizations, and donors- use health information from different perspectives, and for that reason, it is a necessity from the proposed model to fulfil these different requirements.

New advances in ICT provide additional opportunities for utilization of technology to enhance data collection, processing, and exchange; however, developing countries like Sudan is still suffer from a lack of capabilities such as integrated networks. In this model we argue that it is still possible to use existing ICT infrastructure to enhance health information system in Sudan, by using a hybrid data transfer approach including online and offline data communication mechanisms.

B. Governance Organizational Structure – Country (Layer 1: L1)

The Federal Ministry of Health (FMoh) represent the first layer (Layer 1: L1), and governing body which will contains the further sub layers (see Fig. 1). The required data format, and structure will be created in this layer, including the disease coding system and data template structure. In addition, this layer represents the final destination of data after the collection and aggregation processes were completed in the descendent states' layers, to reflect the health status as a national representation format. In addition, the data in this layer could be used for national interventions, analysis, and decision making.

C. States (Layer 2: L2)

The second layer (L2) represents the geo-political division of the country in term of states. Each state in Sudan has a state government ruled by a governor and state ministries. Not all federal ministries have corresponding state ministries; however, all states have ministries of health which are related to the FMoH technically and report medical data to the federal ministry. The data template structure is inherited from the parent country layer (L1) and forms the data definition that is required from each state; furthermore, this definition will be spread downstream to the third layer (L3) for each locality under the corresponding state. The collected data in the state layer reflects the health status in a given state and represents the resultant data aggregation from governing subunits (Localities L3); furthermore, interventions, analysis, and decision making are limited to the state authority only.

D. Localities (Layer 3: L3)

The third layer (L3), the localities, is the organization of governance activities in each state, which is responsible from applying policies and regulations in a specified geographical area and interacting directly with the population and business bodies including health facilities. Although localities are obligated to supervise public health and interventions, they are however, loosely tied to maintaining technical health information. the states' ministries are more related to health providers and facilities, but even in that case, ministries of health in states organize the health data based on localities, and thus this layer can be considered as an organizational data layout at that level. The data template is inherited from the state layer (L2) to form the locality data structure definition. Data in this layer is collected directly from data sources, which are the health facilities (L4), and the aggregated data represents the health status in the specified locality.

At this point, the proposed data model defined how the data should be hosted, structured, collected, and aggregated,

and the data levelling mechanism was introduced; furthermore, the design determines the data usage, authority, and representation. Using a layered hierarchy will provide more flexibility if the data in each layer needed to be compiled with other datasets; for instance, we can compile the model data with a civil registration system, medical insurance data, or budget data, because the national data is structured in the same design, and in that context the model output will enhance future data compilation and integration needs and reduce the time consumption of further data preparation activities.

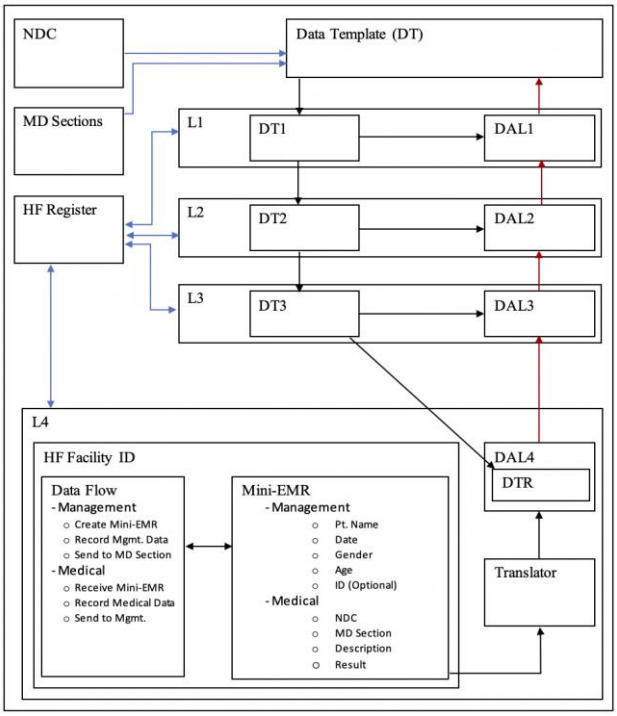

Fig.1: An overview of the Multi-layered Data Attendance Collection Model (MDACM)

Legend:
L: Layer
DT: Data Template
NDC: National Disease Code
DAL: Data Aggregator in Layer x
HF Register: Health Facility Register
Mini-EMR: Minimized Electronic Medical Record

E. Medical Data Management

Data in the medical sector is shared and used by different stakeholders for different purposes and interests, which include: patient medical records, disease information and management data. However, data links, relations, and semantics are always challenging, especially technical health information concerned with diseases. For example, in a health facility, a patient who enters the health facility for clinical consultation may be redirected to a medical laboratory for investigation of a specific disease such as malaria. A positive laboratory result of the requested disease test will support the physician's order for medicine placed with the pharmacy, which will finally be issued by the pharmacist. In this example, three medical sections (clinic, laboratory, pharmacy) interact with the mentioned disease from different perspectives, in particular, the data format and type of that disease; however, the common denominator is malaria, which is processed differently in each section and needed to be mapped and linked in all medical sections, and thus, building a basis for disease codes becomes crucial for disease analysis and control.

The relations between medical data records can be identified using diseases' standard codes, by unification of the language between different medical stakeholders, which facilitates data links and semantics even in different environments and usages; moreover, further data analysis can be enhanced if the data has been properly standardized, referenced, and normalized. This should be considered in early data design stages and thus diseases need to be modelled and standardized.

F. National Disease Code

Top health management is responsible from producing and maintaining National Disease Code (NDC) records for each identified disease in the country. An NDC is a unique numerical or string value that defines a specific disease (see Table 1), and this code will be used to link the data for all data processing activities concerning this disease. In addition, Health facilities will integrate and report their data regarding this disease using the NDC as a key, and furthermore all data templates use the NDC in the template's definition structure to refer to the required disease.
An example of NDC records could look like the following table:

TABLE 1
DEFINED NATIONAL DISEASE CODES (NDCS)

No.	NDC	Disease Name
1	10001	Malaria
2	20001	Cholera
3	30001	Flu

NDCs can be further nested to accommodate many types for a specific disease; however, we are going here to use a simple single disease code in this model.

G. Medical Sections Organization

The health system in a country is constructed and managed by different specialized bodies inside the ministries of health; for instance, in Sudan clinical data are maintained by sub-administration (disease control) in the health ministry, drugs are managed by the National Fund for Medicine, and medical laboratories are supervised by laboratory administration in the ministries of health. On examining medical data for diseases particularly, we can identify the relations between these medical entities concerning a given disease, for instance, diagnosis, laboratory testing, and curing of the same disease, and we notice that data from different entities can be collected and mapped to that disease. This relation can gain significant importance when the data is to be collected at national levels and the integration between different sections can enhance the data analytical process. To create such a relation, we integrate NDCs with designated medical entities (sections) by listing the sections that are involved in each disease (see Table 2); in addition, medical sections are used to link health facilities' internal medical sections to corresponding layered health bodies; for example, pharmacies are linked and report to the drugs and medicines administration.

H. Data Templates

When data is collected from different data sources, it is normally represented in different data formats and structures,

which puts an extra burden on data preparation activities in terms of transforming different datasets from different data sources into a unified data-ready structure to enable data analysis operations. The time and effort consumed in this process always dominate almost every analytical project; however, this problem can be optimized by instantiating pre-defined and constrained data templates and install these templates in data sources before data collection starts.

TABLE 2
INTEGRATION OF NDC IN MEDICAL SECTIONS

No.	NDC	Medical Section
1	10001	Clinic
2	10001	Medical Laboratory
3	20001	Clinic

In this model, the data template implementation mechanism is introduced for the data collection process inherited by data sources. It is designed to constrain the required data format, type, and values in lower stages (data entry points), unlike traditional approaches which collect data and then optimize it later. This mechanism transfers and distributes the burden of data preparation activities to the start point; moreover, data cleansing, transformation, integrity, and constraints are implemented and corrected at the data source at the time of collection, which enhances the collection process in terms of time, effort, and cost.

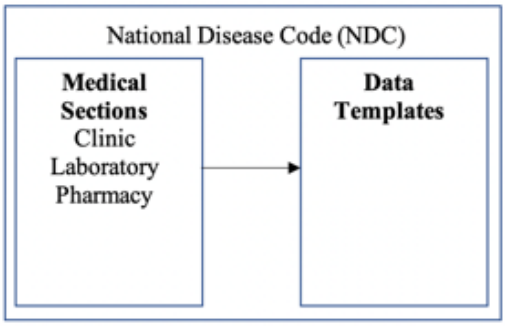

Fig. 2: National Disease Code

At this point, an abstract of the data (template) is created and a definition of the data is structured in terms of format (**how**), reporting dates (**when**), data source (**who**), and the required data (**what**), in addition to injecting the pre-defined reporting layer hierarchy to insure proper data organization (**where**). An instance of a data template represents the reporting of data (medical data) from a specific health facility (data source) at a specific date (attendance of data). To enhance the management process, the data template is organized into two sections, first, the management section, which is concerned with the organization and administration of the data-reporting templates, and secondly the medical section, which focuses on the medical data results and values.

1) Data Template Management Section:

The management section of a data template starts by providing identification of a data source (health facility). As each health facility is registered in the system, the identification will directly lead to mapping the data source location to the corresponding parent layers (locality –> state –> country); moreover, it implements an authentication mechanism (registered and authenticated facility).

Health information analytics and statistics use time and space dimensions in reporting data. The space (geographical area) was already inherited previously by using health facility identification in this model, and the time of data reporting is managed by the reporting-date property in the data template. The reporting date not only acts as a timestamp for the data but can also be used to identify missing (absent) data, in addition to the attendance status property which provides the status of the data template. The following values of attendance status are used:

- Pending: The data template reporting date has not been reached.
- Present: The data template date has been reached and the data has been uploaded.
- Missed: The data template date has been reached and the data has not been uploaded.

Constraints are applied to the data entry time to check for possible data errors, for example, incomplete data, format, translation, and conversion, and correction takes place. Each row in the template is referenced by a data source identifier in addition to the date field, which enables the collection process to update the data row rather than creating a new data record.

The proposed model introduces a mechanism to collect and upload data in term of the date (timestamp) of the data. This mechanism can be used to enhance the collection process from monthly reporting (monthly aggregate data) to daily reporting and even real-time reporting by continuously updating the template's instance-specific date. The collected data is aggregated in the data sources daily, which can optimize the analysis and decision-making process in emergencies and catastrophic events which need instant feedback, for example disease outbreaks.

2) Data Template Medical Section

Medical data is concerned with disease-related information and is organized in this section, first, by identifying the reporting unit (medical section) within the health facility which is the data owner, for example the clinic, laboratory, or pharmacy. Although the medical section will provide additional data aggregation criteria in parent layers, it also enables a data traceability feature by determining which medical unit is responsible and has reported what data; moreover, it enhances the data correction process if needed. The NDC is used to constrain the collected data about a specific disease.

Although the main objective of the model is to properly collect accurate medical information about specific diseases, medical data about patients is usually represented in the form of age groups or ranges, for instance, infants, adults. In addition, the gender also characterizes and classifies the medical representation, especially if the data is required in medical analysis. For that reason, the model structures the values of medical results in two extra levels: age and gender. A matrix organization is formed by adding age groups as rows and gender types as columns. A summation of rows will provide the total number of patients for the data template.

To enable more insight into the data results, optional reference values are attached to the template data; for instance, laboratory results are often supported by a normal range, which can be standard default values and can be

updated by local reference values for the health facility to justify its result values.

Different data sources may use different results formats to report their data. For this purpose, a translation mechanism should be implemented in the data template to map the local data format into unified format and data types; for example, the result can be represented as [Found|Not Found], [Yes|No], [True|false], or [0|1], which all represent Boolean values, and the string results can be coded and mapped to numeric values to make quantitative analysis easier.

The proposed model provides, at this point, the necessary infrastructure for the data collection process by determining how the data will be collected and aggregated; in addition, it provides a constrained data template model for the health facilities.

I. Positioning and Registration of Health Facilities (Layer 4: L4)

Health facilities are registered and licensed to be able to provide healthcare services in Sudan, while the FMoH supervise this process, it is actually practiced by sub-state ministries; as an advantage, the model will use current registration process to identify health facilities. Registration of health facilities is integrated in the model as the last layer (Layer 4: L4) and positioned under the locality layer (L3), where the health facility is physically located. The data in this layer is aggregated locally on a daily basis for each disease and represents the local statistics for a specific health facility.

To properly organize and link any health facility organizational structure with the data model, medical sections in the health facility should be identified and integrated with the corresponding master medical sections in the master layer (L1). This step provides a mechanism to manage subsets of data, for instance the laboratory data for a specific disease. In addition, the medical section in the health facility will inherit a full set of concerning NDCs and the health facility will be able to exchange data with the model.

J. Mini-Electronic Medical Record (Mini-EMR)

In this section, we introduce a minimized version of the Electronic Medical Record (mini-EMR) to address the problems of data collection in health facilities which occur between medical and management staff, who are mutually responsible for recording and reporting health information and statistics in the health facilities. While management personnel experience problems due to incomplete data, missing data, reporting delays, and unavailability in some cases, the medical staff on the other hand, prioritize service provision instead of data management, which as a consequence, affect their data recording commitments, in addition to other challenges such as: lack of medical staff, training skills, and a preference for using the data recording time to provide health care to waiting patients. Apparently achieving a balance between the provision of healthcare and strengthening of the health information system is a difficult trade-off and an intermediate solution should be introduced with continuous optimization. Medical staff are obligated to record clinical information while management staff are responsible for aggregate medical and management information, and from that perspective, a simplified and structured medical patient record with a data flow model can enhance the data collection process by reducing the number of fields in the EMR to create a smaller version. The mini-EMR is structured into two sections: the first contains the patient's personal information and management data and the second contains the medical data. In addition to a simple process flow between the two sections, the mini-EMR addresses and adopts the model organization and criteria by incorporating data links and interfaces for the data aggregation process such as dates, the medical sections, and the NDCs. An example of a mini-EMR could be introduced as follows:

Mini-EMR

Data Structure

- Management Section
  I. Patient name
  II. Gender
  III. Age
  IV. ID (optional)
  V. Date
  VI. Health facility ID
- Medical Section
  I. Medical section name (clinic, laboratory, pharmacy, etc.)
  II. NDC
  III. Description (clinical data, test name, medicine name, etc.)
  IV. Result (diagnosis, laboratory result, drug dose, etc.)

Data Flow

- Management
  I. Create mini-EMR
  II. Record management section data
  III. Send to medical section
- Medical
  I. Receive new mini-EMR
  II. Record medical section data
  III. Send for management and aggregation
- The minimized EMR can be used as a starting point for implementing a full and comprehensive electronic medical record in the future or according to the development of the health facilities toward full ICT adoption in the health information system; however, the proposed structure is used to enable data aggregation from health facilities (DAL4) while at the same time reducing the data collection problems between medical and management staff and introducing the data sharing mechanism inside the health facility in compliance with the general data model requirements.
- Data aggregation in the health facility layer (DAL4) uses the medical section data to aggregate the data into the data template row (DTR) by transforming and aggregating the patient records into a disease aggregation record using mini-EMR parameters such as date, NDC, gender, and results.

K. Data Configuration

In order to start the data collection process, the system should be initialized and configured. The configuration process starts by determining the target collection period, for instance, annual collection, which constrains the model by start and end dates. The next step is selecting the required data template from the templates list, which identifies which disease is targeted and in which medical section; for example,

the objective may be to target malaria disease test results in laboratories annually. The model generates data template rows for all dates for the collection period, and default values are initialized for quantitative attributes, in addition to performance indicators such as the attendance status.

The data template, at this point, was created in the master layer (L1) and instantiated subsequently in Layers 2 (States), Layer 3 (Localities), and Layer 4 (Health Facilities) to produce instances of datasets.

L. Data Transaction

Datasets are collected from health facilities according to referenced data template. Each registered health facility uses a combination of a health facility identifier and reporting date to authenticate its reporting data. The complexity of the model structure is simplified in the health facilities by mapping the registration ID to the entire structure (see Figure 3), and the main focus of the health facility is to accurately collect and report the required medical data. The data aggregator in the fourth level (DAL4) is responsible for aggregating (summation, counts) the health facility's data by using date attributes. The process is organized as follows:

- Collect and aggregate data in the health facility (DAL4) daily.
- Use the date key to match data to the DTR and the health registration ID to identify the health facility.
- Identify missing and incomplete data.
- Correct data errors and assign default values if needed.
- Translate the result type to match the template codes.
- Provide the results to the user and ask for confirmation.
- Update the DTR result data and attendance status.

M. Data Aggregation in Organizational Layers

Ready datasets from health facilities are used to generate upper layers' datasets, in particular, three additional data aggregators (DAs) are generated as follows:

- Data is aggregated from all health facilities in a locality and the attendance performance is assigned in the localities layer (DAL3).
- Data is aggregated from all localities in a state and the attendance performance is assigned in the states layer (DAL2).
- Data is aggregated from all states and the attendance performance is assigned in the country layer (DAL1).

At this point, the model provides a structured dataset for a specific disease with the ability to verify, validate, and trace the data to the data sources; moreover, the datasets are represented in multiple dimensions, which are:

- Geographical distribution (space dimension): provided by the layered structure;
- Periodical distribution (time dimension): provided by articulating the data template by date ranges;
- Disease-specific dimension: provided by standardizing and linking the reported data using the NDC;
- Gender distribution: provided by structuring the dataset values by gender;
- Age distribution: provided by using the age group categorization for datasets.

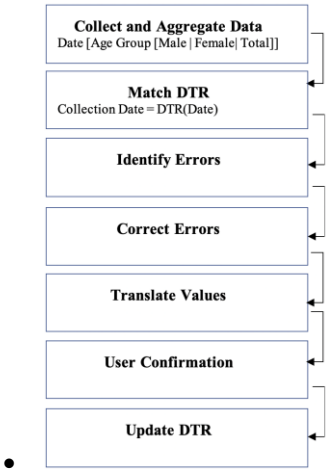

Fig. 3: Data Transaction Legend: DTR: Data Template Row

V. Discussion

Developing countries -including Sudan- are making significant steps in adapting ICT in governance; however, many challenges are still outstanding, which can be found in the infrastructure, cost, and capacities. The health sector always has the dilemma of assigning the budget between health care provision and development to improve data collection, quality, and decision making. Our model provides an optimization, in that regard, by addressing these challenges with respect to data preparation issues as well. The model provides a comprehensive solution for health data collection in the country with respect to the global consensus and guides of health information systems, for example managing the disease code system and structuring records for health facilities and patient data. Many approaches to national data management focus on patient records like the Aadhaar card in India and Côte d'Ivoire [25]; however, these models which focus on patient records rather than the health service providers – the health facilities – may suffer from data analysis problems if it is necessary to reflect the country's geo-political status, because of instability of the moving human patient versus the fixed health facilities, which can have a significant impact on health planning, resource management, and decision making. The ability to manage diseases, using the NDC for example, can help decision makers and health programmes in a country, by providing a mechanism to monitor all diseases in one national data warehouse with the ability to split datasets or even entire data marts at the same time as focusing on a special disease or health programme.

Another contribution of the proposed model, it combines different analysis properties early in the model design; for instance, the timeline, gender, and group ages. By considering such important parameters in the data gathering phase, the analysis process can be enhanced with rich datasets that could be analysed with practical health indicators combined with geographical distribution. In addition, the model is prepared to manage historical and archival data using dates and attendance data, which can improve decision quality and data-pattern discoveries.

As many researchers point to the need for data quality in big data management [1,15], developing data models that can adopt data quality measures and practices to strengthen health information systems is becoming an outstanding challenge. In this regard, the model provides a mechanism to adopt such a

recommendation in data preparation by distribution of the workload at different levels, in particular the ETL process, which can improve the processing speed, data quality, and error correction.

Application of the model is flexible from the data exchange point of view. The data can be transferred using different methods including online and offline connections, depending on the availability of networks and communications capabilities in the country. Configuration of data templates with dates can allocate data on arrival even in the situation of unavailable connectivity, and thus, overcome the necessity for a pre-existence of communication for immediate data exchange, which can offer an aspect of cost reduction. However, in good economic situations, the model can be further improved using fast online networks without the need to make significant changes.

-Model Implementation

Implementation of the proposed model should systematically follow a top-down setup approach. First, the health authority in the country is requested to provide the required framework, policies, and procedures in order to set up the health information system for the country, for example Sudan. Setting up the system includes, definition of the geo-political structure of the country by dividing the country into matching states and localities. A tree view hierarchy will start to emerge, which provides a visual layout of the upcoming data and represents geographical areas with political governance.

Secondly, the health authority should keep a national registration records for health facilities in the country, and each health facility should be allocated using the geo-political organization that was defined previously. The allocation should determine the geographical position of the health facility inside the locality layer and state layer respectively. This step will enable health managers to visualize the health capacity and provision status in the structured area. Not only is this useful to organize health data; it will also provide the benefit of giving insights about health resource allocation, gaps, and needs.

At this point, the framework was implemented with four layers: country (FMoH), states, localities, and health facilities. At the bottom the data sources (health facilities) were identified and defined as data collection points; moving up the hierarchy, three data aggregation levels were identified. Moreover, data links, flow, interfaces, and traces paths were clearly visible.

The third step is to define a standard disease coding system, in particular, the NDC. The health authority in the country should create and code a unique list of diseases in the country in order to create a foundation for data links for health data. The NDC will be the reference key that will be instantiated, shared, and inherited between all stakeholders in the system. Master medical sections should be created to match different medical sections in health facilities, for example, laboratory, medicine, and clinics (disease diagnosis). Of course, health management has specialized administration and subsections that manage each medical domain; however, a virtual representation of theses sections will simplify data classification in the upcoming data management activities; moreover, it will map the actual medical section with the health facility directly to the collection model and hide the actual complex distribution of the real medical administration.

In the fourth step, to start using the system, a data template should be created at the health authority level (Layer 1). The data template creation will define what disease is targeted by using NDC for the data collection process, and this will automatically identify the targeted medical sections which is previously configured in the master medical section records; in addition, the data template will contain the time duration of the targeted data, empty result fields, and data attendance status flags. Data is structured in the data template to represent different data criteria such as patient's age intervals, gender, and result transformation references. The creation of a data template enables health administrators to define their expectations of the model output in addition to data rules and constraints. The quality management measures can be created and applied, and in addition, the data template illustrates (Layer 1) the expected output shape of the data.

In the second layer (the state layer), the data template should be instantiated by the number of states. In the context of Sudan, an 18 layers (L 2) data templates will be created. Each targets one state. In the same context, each Layer 2 data template (state template) will be further instantiated by the number of localities in the designated state, and thus the Layer 3 data templates are created (locality layer). Each health facility should obtain a copy of the descendant Layer 3 data template, which defines the data required from the health facility and its format and constraints. The data template for the health facility is linked with the health facilities register and can be used as an authentication mechanism to identify the source of the data. At this point, the data templates in the model are created and sent to data sources, to be filled by health data.

Adoption of the system depends on the available ICT infrastructure and available resources, in order to work with optimal performance. At the health facility, a daily aggregation of the data for a specific disease will form the output for a specific record in the health facility's data template – the DTR. However, it can be flexible in order to accept many forms of collection and aggregation mechanisms for health data depending on the currently available ICT resources at each health facility. For instance, the following scenarios may occur:

- the health facility has an electronic HIS and the system can be directly integrated with the DTR;
- the health facility adopts the Mini-EMR to collect and aggregate data and upload it to the DTR;
- the health facility uses manual records, performs manual aggregation, and enters the output in the DTR;
- the health facility has no ICT resources and manually aggregated data is provided to the locality layer;
- the locality layer has no ICT resources and manually aggregated data is provided to the state layer;
- the state layer has no ICT resources and manually aggregated data is provided to the country layer.

The data is exchanged between multiple layers in bi-directional mode. On the creation of the data template in the first layer (L1), instances of the state layer are created and sent down the hierarchy to the states (L2). In the same fashion, instances of data in each L2 are created and sent to L3. Each locality creates and sends data templates to all health facilities in that locality. The exchange of data can benefit from the available network connectivity to automate

the data transfer and provide online data exchange; however, in the case of unavailability of network coverage, the instances are created in the corresponding layer and maintained in that layer. In addition, in the case of disconnection of the network, the system can operate in offline mode and should provide a synchronization mechanism to upload or download the data when restoring the connection. Depending on the available resources and technologies, the system implementation can adopt many other forms of data exchange, for example, cloud computing and shared storage; however, in limited-resources environments, additional extracting and loading tools could be developed; for instance, a tool can be provided to extract and upload the template data stored in spreadsheets or other file formats.

Data quality management should be considered in the design of the data templates in the first layer, and quality rules and constraints should also be defined and equipped with the data template. This will enable the system administrators to implement shared quality policies early, to minimize the burden of the data preparation load. When collecting the data from health facilities, translation rules should be applied so that the results match the required output format of the model. Next, the data aggregator is implemented inside the health facility – the data aggregator Layer 4 (DAL4) – and the quality rules are checked. In the case of data errors, the system should provide a repair mechanism or report the errors to the user for correction.

The system will continue to aggregate the data in a backward (folding) fashion from localities (DAL3: from all health facilities in the locality) to states (DAL2: from all localities to states) to country level (DAL1: from all states to the health authority, FMoH). The aggregated data in all levels should maintain quality status measures which report the level of availability (attendance), accuracy, and error ratios in order to determine the degree of reliability of the collected data. After the process has been completed, the model will provide hierarchal datasets that match the organizational structure (geo-political), which will be ready for analysis and usage.

Datasets that will be produced from the data model can have multiple usages and advantages. For instance, they can show the status of specific disease in the country, it can be used to analyse specific interventions for a disease, and moreover, it can enhance the allocation and management of resources and of course increase the visibility of the decision-making process.

The implementation of MDACM can have many shapes depending on the ICT resources available and upgrades, moreover, a phase-based implementation can take place without the need for significant modification of the data model. In addition, the model can also operate in a limited-resources environment; however, the model will continue use and enhance the quality measures of data in the health information system and ensure proper data representation of the health status in the country by considering the ICT resources and availability – even if limited – as an interim process toward health information system optimization rather than as an impossible obstacle.

The implementation of electronic health information systems at national scale is directly affected by the country's financial resources and capacities, which leads to an observable distinction between lower and higher income countries. As a consequence, many system adoption models [27] have been introduced to systematically advance the process of health system automation in developed and developing countries. One solution to minimize the time delay in this process is to use flexible data models, like the one which is introduced in this article. Besides providing a structured mechanism toward the adoption of e-health in developing countries like Sudan, the model contributes by facilitating the transformation process between medical patient records and aggregated data: while the former are required to manage patients' medical treatment and are used by medical and management staff in health facilities, the aggregated data is mainly used for public health management, resource management, and decision-making processes. Our model provides a solution to enhance health information system by: (1) structuring an effective e-health system at country scale, (2) it provides a mechanism to transform patient medical data into aggregated data for upper management, (3) the model facilitates the evolution of an adoption mechanism for an e-health system with consideration of the implementation cost, infrastructure, and capacities of lower income countries, (4) it provides a starting point for data quality management to enhance the quality of collected data in the health system in the country, and lastly,(5) it provides a mechanism to comply with international health recommendations like standardization of disease codes and profiling health facilities in the country[24].

-Example

The example demonstrates the applicability of the proposed data model. The simulation uses MCDAM for collecting and analysing the data for the large Malaria outbreak in Sudan during 01-Jan-2019 to 31-Mar-2019. The model targets all states and is based on collected laboratory medical results.

First a new data template is created and template instances are generated for all states. This is done in the first layer (L1) (see fig.4). The empty data templates are then sub-instantiated in the next layer (L2, L3 and L4). The actual data collection process starts from the health facilities, using the proposed Mini-EMR model (see fig.5), where the data is aggregated daily for each health facility (see fig. 6). After collecting the data from the health facilities, it is aggregated in a similar way from the health facilities in the locality layer (fig. 7), thereafter in the state layer (fig. 8) and finally in the master layer (fig. 9).

In this simulation, a data status field is attached representing the data attendance. The successful data collection for all specified period is indicated by a "Present" status. The contextual, partial collection is indicated by "Partial Completed" stating also the last date of the data collection (Data Date). Finally, absence of data is designated an "absent" status.

VI. Conclusions

In the article, we tackled the manual data collection problem in the health sector in Sudan. Many challenges were discussed including disease coding models, data collection processes, data preparation, and data quality. The proposed model provides a mechanism to collect health data in a multi-layered fashion that represents the geo-political layout of the country and thus enables data compilation with similar governing structures that were found at national levels. Moreover, it provides an efficient and structured data

Fig. 4: Data Template for Malaria Disease for All States in Sudan from 01/01/2013 to 31/012019

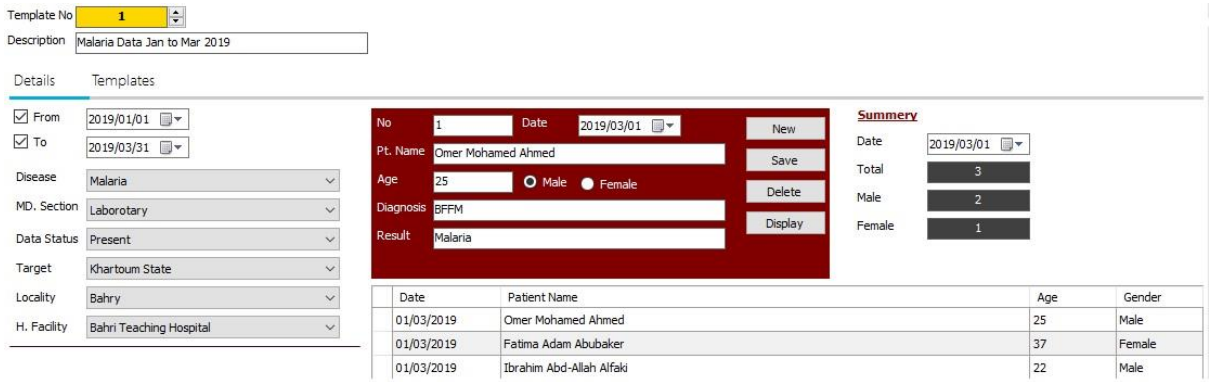

Fig. 5: Use Mini-EMR to Collect Data for Malaria Disease Khartoum State-Bahry Locality- Bahri Teaching Hospital in Sudan for 01/03/2019

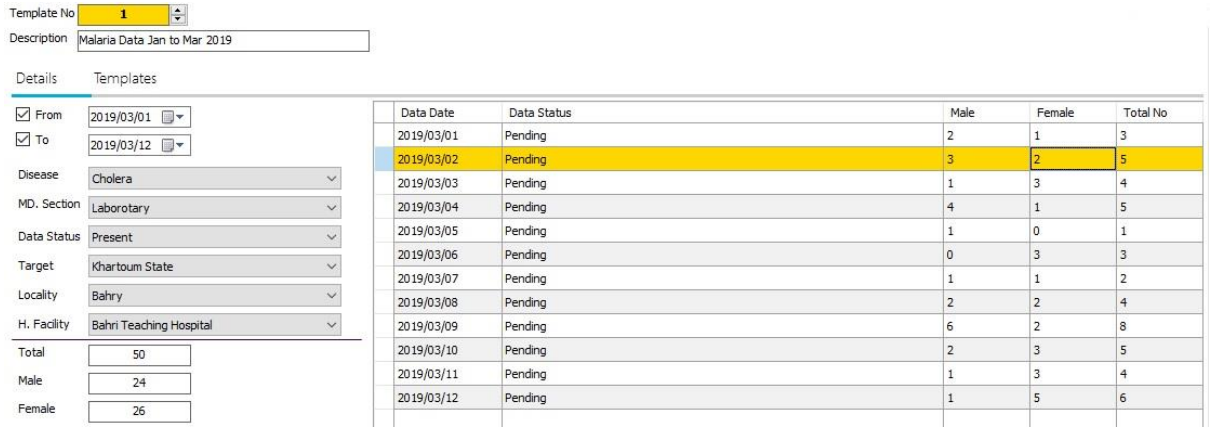

Fig. 6: Aggregated Data for Malaria Disease Khartoum State-Bahry Locality- Bahri Teaching Hospital in Sudan from 01/03/2019 to 12/03/2019

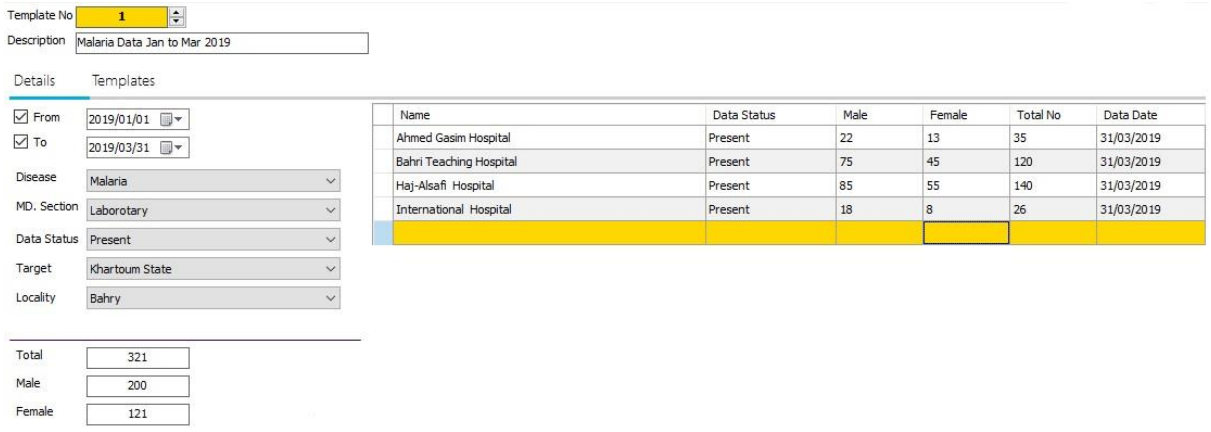

Fig. 7 Collected Data Template for Malaria Disease Khartoum State-Bahry Locality in Sudan from 01/01/2019 to 31/03/2019

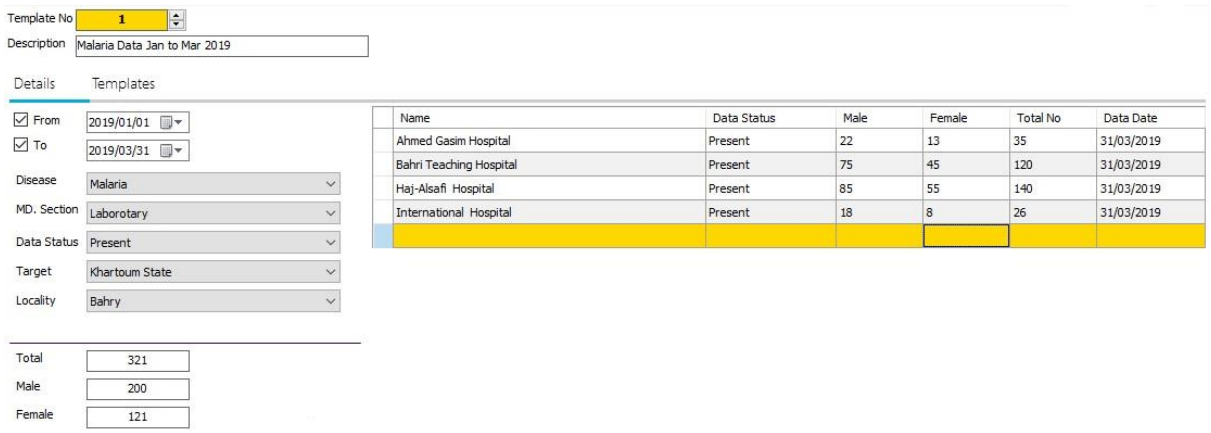

Fig. 8 Collected Data Template for Malaria Disease Khartoum State in Sudan from 01/01/2019 to 31/03/2019

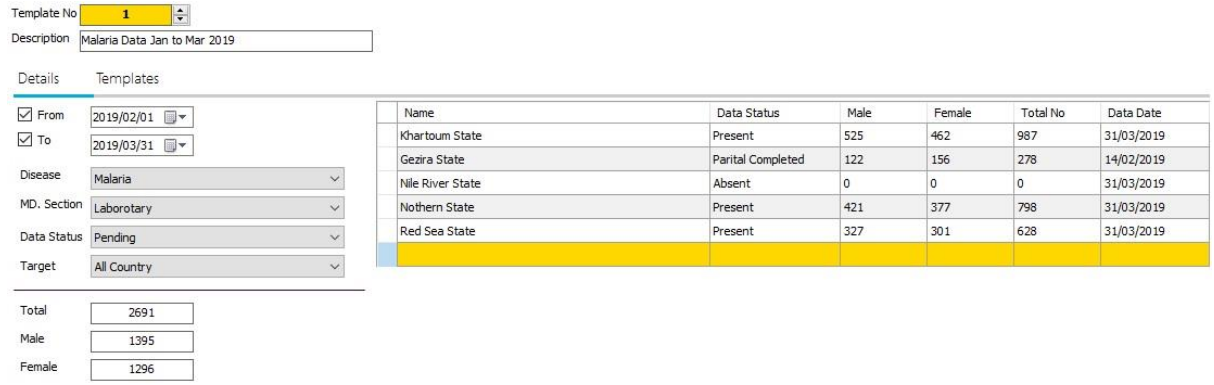

Fig. 9 Collected Data Template for Malaria Disease for Selected States in Sudan from 01/01/2019 to 31/03/2019

collection method with implementation of data quality and preparation measures to minimize the processing efforts, cost, and time consumption in this regard. The concept of measuring the attendance data is used to detect the status of the data early on and thus to improve data analytics and decision making.

REFERENCES

[1] R. Dubey, A. Gunasekaran, S. J. Childe, S.F .Wamba, T. Papadopoulos, "*The impact of big data on world-class sustainable manufacturing"*, vol.84, pp 631-645, 2016.

[2] G. Walt, J. Shiffman, H. Schneider, S. F. Murray, R. Brugha, L. Gilson, "*Doing health policy analysis: methodological and conceptual reflections and challenges, Health Policy and Planning"*, vol. 23, 2008.

[3] H. Busse, E. A. Aboneh and G. Tefera, "*Learning from developing countries in strengthening health systems: an evaluation of personal and professional impact among global health volunteers at Addis Ababa University's Tikur Anbessa Specialized Hospital (Ethiopia)"*, 2014.

[4] D. J. Casley, D. A. Lury, "*Data collection in developing countries"*, 1980.

[5] A. Skoogh, Anders and J. Björn, "*Time-consumption analysis of input data activities in discrete event simulation project"*, vol.1, 2007.

[6] R. Ahmed, R. Robinson, A. Elsony, R. Thomson, S. B. Squire, R. Malmborg, P. Burney, … K. Mortimer, "*A comparison of smartphone and paper data-collection tools in the Burden of Obstructive Lung Disease (BOLD) study in Gezira state, Sudan. PloS one*, 2018

[7] M. N. Sarkies, K. A. Bowles, E. H. Skinner, D. Mitchell, R. Haas, M. Ho, K. Salter, K. May, D. Markham, L. O'Brien, S. Plumb, T. P. Haines, "*Data collection methods in health services research: hospital length of stay and discharge destination. Applied Clinical Informatics"*, vol. 6(1), pp 96-109, 2015.

[8] T.C. Redman, Data Quality for the Information Age, Norwood-USA MA: Artech House, 1996.

[9] W.W. Eckerson, "Data Quality and the Bottom Line: Achieving Business Success through a Commitment to High Quality Data",2002

[10] Trifacta, *Global Organizations Wasting Billions of Dollars on Data Preparation.* [Online]. Available : https://globenewswire.com/news-release/2018/05/17/1508217/0/en/Global-Organizations-Wasting-Billions-of-Dollars-on-Data-Preparation.html, accessed January 2019

[11] S. D. Yawson, G. Ellingsen, "*Assessing and Improving EHRs Data Quality through a Socio-technical Approach"*, in Procedia Computer Science, vol. 98, pp 243-250, 2016,

[12] E. Rahm, H. H. Do, "*Data Cleaning: Problems and Current Approaches"*, 2000. [Online]. Available: http://dbs.uni-leipzig.de

[13] S. Zhang, C. Zhang, "*Data Preparation for Data Mining"*, vol. 17, pp 375-381, 2003

[14] Kwak, S. K., & Kim, J. H. Statistical data preparation: management of missing values and outliers. Korean journal of anesthesiology, 70(4), 407–411. doi:10.4097/kjae.2017.70.4.407,2017

[15] Aguinis, H., Hill, N. S., & Bailey, J. R. Best Practices in Data Collection and Preparation: Recommendations for Reviewers, Editors, and Authors. Organizational Research Methods. https://doi.org/10.1177/1094428119836485,2019

[16] S. Kanchi, S. Sandilya, S. Ramkrishna, S. Manjrekar and A. Vhadgar, "*Challenges and Solutions in Big Data Management",* 3rd International Conference on Future Internet of Things and Cloud, Rome, pp. 418–426, 2015

[17] L. Omran, V.C. Storey and R. Y. Wang, "*Systems approaches to improve data quality"*, 1995

[18] . Kristian, R. Filipe,C. P. Francisco, Mobility Patterns, Big Data and Transport Analytics, Pages 73-106, ISBN 9780128129708,https://doi.org/10.1016/B978-0-12-812970-8.00005-1, 2019

[19] Heinrich, Bernd, Kaiser, Marcus and Klier, Mathias, "*How to measure Data Quality? A Metric-based Approach"*, 28th International

Conference on Information Systems (ICIS), Queen's University Montreal, Canada, 2007.

[20] C. Cappiello, C. Cerletti, C. Fratto, and B. Pernici, "*Validating Data Quality Actions in Scoring Processes",* Journal of Data and Information Quality, vol. 9, 2018.

[21] M. D. Angeles and G. U. Francisco, "*A Data Quality Practical Approach"*, vol., pp 259-274, 2009.

[22] Svetlozar N., J. Boris, J. Nenad, S. Abhishek, R. Sippo Rossi,, Generating insights through data preparation, visualization, and analysis: Framework for combining clustering and data visualization techniques for low-cardinality sequential data, vol 125, https://doi.org/10.1016/j.dss.2019.113119.2019

[23] Konstantinou N., Paton N. W. Paton. Feedback driven improvement of data preparation pipelines, 2019

[24] (2017) World Health Organization, *Health Systems*. [Online]. Available: http://www.who.int/topics/health_systems/en/

https://doi.org/10.1016/j.is.2019.101480.

[25] R. Wyber, S. Vaillancourt, W. Perry, P. Mannava, T. Folaranmi and L.A. Celi, "*Big data in global health: improving health in low- and middle-income countries"*, vol. 93, pp 203-208, 2015

[26] World Health Organization, *Health Metrics Network Framework and Standards for Country Health Information Systems*, January 2008. [Online]. Available: https://www.who.int/healthinfo/country_monitoring_evaluation/who-hmn-framework-standards-chi.pdf

[27] (2012) World Health Organization, *Management of patient information trends and challenges in member states*. [Online]. Available: https://apps.who.int/iris/bitstream/handle/10665/76794/9789241504645_eng.pdf;jsessionid=A98788C197B17146B2C795E57D4C89CB?sequence=1

[28] J.T. Bram, B. Warwick-Clark, E. Obeysekare, and K. Mehta, "*Utilization and Monetization of Healthcare Data in Developing Countries"*, pp. 59–66, DOI: 10.1089/big.2014.0053,2 015.

[29] (2017) Human Rights & Health Equity Office, *Guide to demographic data collection in health-care settings, Sinai Health System.* [Online]. Available: http://torontohealthequity.ca/wp-content/uploads/2017/10/Measuring-Health-Equity-Guide-to-Demographic-Data-Collection.pdf

[30] A. E. Monge, "*Matching Algorithms within a Duplicate Detection System"*, 2000.

[31] Clinical Data Classification. [Online]. Available: http://guides.lib.uw.edu/hsl/data/findclin

[32] R. Pelánek, J. Rˇihák, J. Papoušek, "*Impact of Data Collection on Interpretation and Evaluation of Student Models"*, he Sixth International Conference on Learning Analytics & Knowledge, pp. 40-49,2015.

[33] (2007)The World Bank, *The World Bank Strategy for HNP Results Annex*. [Online]. Available: http://documents.worldbank.org/curated/en/102281468140385647/Healthy-Development-the-World-Bank-strategy-for-health-nutrition-population-results

[34] K. Bhalla, J.E. Harrison, S. Shahraz, L.A. Fingerhut, "*Availability and quality of cause-of-death data for estimating the global burden of injuries"*, vol. 88, pp. 831-838, 2010.

[35] A.D. Black, J. Car, C. Pagliari, C. Anandan, K. Cresswell, T. Bokun, "*The Impact of eHealth on the Quality and Safety of Health Care: A Systematic Overview"*, PLoS Med 8(1): e1000387. https://doi.org/10.1371/journal.pmed.1000387, 2011.

[36] A.E. Powell, H.T.O. Davies, R.G. Thomson, *"Using routine comparative data to assess the quality of health care: understanding and avoiding common pitfalls"*, BMJ Quality & Safety, 12:122–128, 2003.

[37] J. Thuma, "*Practical Approaches to Data Quality Management in Business Intelligence and Performance Management"*, 2009.

[38] World Health Organization, *Creating a Master Health Facility List*. [Online]. Available: https://www.who.int/healthinfo/systems/WHO_CreatingMFL_draft.pdf

[39] M. C. Azubuike and J. Ehiri, "*Health information systems in developing countries: Benefits, problems, and prospects"*. The Journal of the Royal Society for the Promotion of Health, vol. 119, 1999.

[40] A. M. Abd-Alrhman, L. Ekenberg, "*Modelling health information systems during catastrophic events – a Disaster management system in Sudan", 2017 IST-Africa Week Conference (IST-Africa)*, Windhoek, pp. 1-9. doi: 10.23919/ISTAFRICA.2017.8102390, 2017.