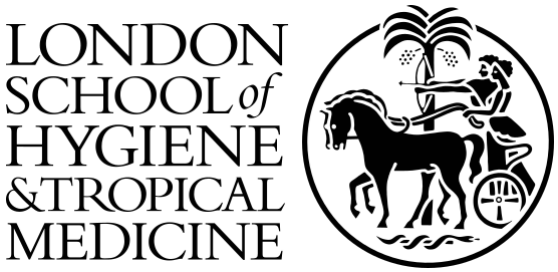
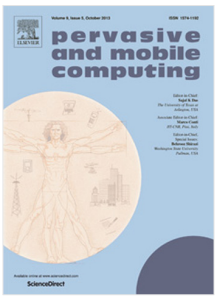


Accepted Manuscript



Using big data analytics to extract disease surveillance information from point of care diagnostic machines

Pouria Amirian, Francois van Loggerenberg, Trudie Lang,
Arthur Thomas, Rosanna Peeling, Anahid Basiri, Steven Goodman

PII: S1574-1192(17)30314-0
DOI: http://dx.doi.org/10.1016/j.pmcj.2017.06.013
Reference: PMCJ 853

To appear in: *Pervasive and Mobile Computing*

Please cite this article as: P. Amirian, F.v. Loggerenberg, T. Lang, A. Thomas, R. Peeling, A. Basiri, S. Goodman, Using big data analytics to extract disease surveillance information from point of care diagnostic machines, *Pervasive and Mobile Computing* (2017), http://dx.doi.org/10.1016/j.pmcj.2017.06.013

Using Big Data Analytics to Extract Disease Surveillance Information from Point of Care Diagnostic Machines

Pouria Amirian[1,2], Francois van Loggerenberg[1],Trudie Lang[1], Arthur Thomas[3], Rosanna Peeling[4], Anahid Basiri[5], and Steven Goodman[6]

[1]The Global Health Network, University of Oxford, Oxford, UK

[2] The Ordnance Survey of Great Britain, Southampton, UK

[3] The Oxford Internet Institute, University of Oxford, Oxford, UK

[4] London School of Hygiene and Tropical Medicine, London, UK

[5] Geography and Environment Department, University of Southampton, UK

[6] Medicine and of Health Research & Policy, Stanford University, US

Abstract

This paper explains a novel approach for knowledge discovery from data generated by Point of Care (POC) devices. A very important element of this type of knowledge extraction is that the POC generated data would never be identifiable, thereby protecting the rights and the anonymity of the individual, whilst still allowing for vital population-level evidence to be obtained. This paper also reveals a real-world implementation of the novel approach in a big data analytics system. Using Internet of Things (IoT) enabled POC devices and the big data analytics system, the data can be collected, stored, and analyzed in batch and real-time modes to provide a detailed picture of a healthcare system as well to identify high-risk populations and their locations. In addition, the system offers benefits to national health authorities in forms of optimized resource allocation (from allocating consumables to finding the best location for new labs) thus supports efficient and timely decision-making processes.

Keywords: Point of Care; Big Data Analytics; Internet of Things; Global Health; Machine Generated Data; Machine Learning

1 Introduction

Diagnostic Point of Care (POC) devices are important tools in the battle against infectious diseases as well as other acute and chronic diseases. POC tests can usually run faster than conventional laboratory testing and need less equipment [1]–[3]. Combining the test results data (generated by POC) with patient demographic data results in a comprehensive dataset which can be used efficiently to extract fine-grained surveillance information at individual-

level as well as at population-level. **With the availability of a comprehensive dataset, performing almost any sort of analytics is feasible. Recently various types of real-time health monitoring systems have revolutionized the collection, sharing, and utilizing of data for personalized healthcare. The availability of a comprehensive dataset, advancement in analytical methods and technologies, especially the use of mobile devices, has led to the concept of mHealth** [4]–[6]**. The Executive Board of WHO in 2016 considered mHealth as an important resource for individual and public health services delivery, especially monitoring the health status of each patient. However, there are several barriers to using concepts like mHelath in low and middle-income countries.**

For myriad political, social, privacy, technical, and security issues, especially in low and middle-income countries [7], coupling demographic data at the individual level is very difficult (if not impossible) and needs legal and ethical approval at various levels [8], [9]. If demographic data are decoupled from test results, neither individual-level nor population-level information can be inferred from the test results. **As an example, although access to mobile technologies has rapidly expanded in low and middle-income countries, unclear healthcare system responsibilities, unreliable infrastructure, and lack of consistent data challenge their implementation** [10]**.**

The contribution of this research paper is a new type of knowledge discovery and extraction based on just machine generated data e.g. using IoT-enabled POC devices. In other words, this paper proposes a novel approach for extracting meaningful and valuable population-level insights using just POC generated data including test results, duration of the tests, the location of the devices, errors and warnings, and quality control parameters. A very important element of this type of knowledge extraction is that the POC generated data would never be identifiable, thereby protecting the rights and the anonymity of the individual, whilst still allowing for vital population-level evidence to be obtained.

Along with this contribution, we also explain a rather sophisticated architecture for implementation of a big data solution to support the main contribution. The implemented system of this paper is capable of storing, processing and analyzing the vast amount of data in real-time and batch modes. **This paper uncovers the potential of using big data analytics in the healthcare domain to find useful information in highly valuable (but untapped) POC generated data**. As illustrated in the paper, the implemented system can also provide the POC data to external analytics clients for performing further knowledge extraction using visual analytics, spatial analytics, and advanced analytics.

By using this novel approach, integration of data from heterogeneous POC devices (various devices which generate data in different data structures and encoding) is significantly less complex and time-consuming than existing approaches. In addition, real-time data collection and real-time analytics can be achieved using the approach of this research which supports efficient and timely decision-making processes. For example, the system described in this paper is capable of detecting anomalies automatically in real-time (without the need for human

intervention) to take appropriate actions using alerts and triggers (informing responsible people and/or authorities for example).

In addition, the big data analytics system uses Internet of Things (IoT) enabled POC devices to automate data generation and storage. As is illustrated in the paper, the resulting system has real-world application for extracting valuable and useful information at a population-level from the machine-generated POC diagnostic and laboratory data. The paper has been organized as follows: Section 2 describes the importance of the POC devices. Then the technical requirements of POC are discussed in section 3. Section 4, illustrates the data that is generated by POC devices and then introduces the accessibility issue of the data generated by POC machines. The proposed solution, including the common data structure and data analytics, is explained in section 5. The big data architecture of the proposed solution, its implementation details and the benefits of it over existing approaches, are described in section 6 and 7 respectively. Some important results of descriptive, diagnostics, predictive and prescriptive analytics of the system are illustrated in section 8 and, finally, section 9 concludes the paper by discussing proposed future directions for the research project.

2 The Importance of POC

The POC diagnostic tests are very important in the battle against infectious diseases as well as other acute and chronic diseases. POC tests can usually run faster than conventional laboratory testing, and with less equipment, so they can extend healthcare availability into the community and reduce the number of patients lost to follow-up (LTFU), or the number of treatments initiated too late [11]–[13]. This aspect of POC diagnostics tests is highly valuable for many high-burden infectious diseases such as HIV, TB, malaria and others, where earlier diagnosis and treatment can also mean the difference between life and death [14], [15]. Most studies have concluded that POC devices are suitable alternatives to traditional laboratory devices [11].

Accurate and timely diagnosis of patients has been a key aspect of the response to infectious diseases, especially measures to prevent onward transmission. In the recent Ebola outbreak, the process for differentiating those who have Ebola from those who do not has posed a great danger to patients. During the height of the epidemic, laboratory and surveillance professionals were overwhelmed, which often led to long wait times and caused the patients a great deal of anxiety [16]. Even after samples are collected, they had to be transported to a laboratory with the capacity to perform the complex and time-consuming tests required. In the case of Ebola, the time from sample collection to receipt of result reported to be greater than six days on average [17]. These delays present patients with an unbearable wait and, more importantly, put uninfected individuals at risk of being infected [18], as well as reducing the number of LTFU or the number of treatments initiated late [11]. In this case, the use of POC tests can mean that the tests can be done where the patients are [15]–[18]. In other words, POC devices can be used efficiently to reduce the delay between a patient's arrival at the clinic and a confirmed diagnosis. In addition, it is sometimes possible to use portable (mobile) POC devices and, therefore, eliminate the need for transport of samples and more importantly it is possible to test patients closer to their community. This is very important since most of the population, especially in rural areas particularly in low and middle-income countries, are tested only when

they have access to a close healthcare services [1], [15]. The availability of POC (especially portable devices) can expand the reach of healthcare beyond what a conventional laboratory could do on its own.

3 Technical Requirements of POC

The WHO defined an ideal POC with 'ASSURED' characteristics, which stands for Affordability, Sensitivity, Specificity, User-friendliness, Rapid results, Equipment-free and Delivered. Based on this definition an ideal POC would bring the test to the patient in an expedient and timely manner. Based on the WHO's vision working with ideal POC requires little technical training or administrating or interpreting. In practice, very few POC diagnostics devices meet all of the ASSURED criteria[2], [16], [23]. Critical results (like positive drug resistant TB or CD4 t-cell counts of less than 200 in HIV infection) of some POC tests need to be confirmed by conventional laboratory tests. In addition, some POC platforms were designed for use in specific laboratory settings and take several hours to run the tests. In other words, they are hardly meet the "rapid" characteristic. However, in some cases, even the above-mentioned devices have revolutionized the availability of rapid, accurate diagnosis of some serious diseases (especially drug-resistant TB) [24]. Some POC diagnostics, including HIV and CD4 rapid tests, are readily available and highly transportable. As an overall statistics, in 2013, 58 million people were tested using HIV rapid tests [25].

From a technical point of view, although the POC devices are very effective, their full potential is limited to their connectivity features and the environment in which they are used. In order to extend the usefulness of POC devices, two important technical challenges need to be addressed: connectivity of devices and analytics of the machine-generated data.

The connectivity of devices means the POC devices need to be connected to a communication infrastructure (wired or wireless network) in order to upload data to databases at both the local-level (city, region or state) and the national-level. The connectivity allows control programs to monitor the quality of tests and testing, and optimize supply chain management; thus, increasing the efficiency of healthcare systems and improving patient outcomes [23].

Most hospitals and clinics rely on laboratories (external or internal) for test results. In other words, laboratories have the POC devices and the actual tests are run in laboratories after samples are received from clinics or hospitals. In most low and middle-income countries, there is no sufficient digital network infrastructure or regulations to send the results of tests back to the clinics and hospitals electronically. Apart from regulations, there must be an automatic mechanism to send the test result data back to clinics and hospitals in order to record test results along with patient data (local-level connectivity). Also, a centralized database should be hosted by Ministry of Health (MOH) or any appropriate national-level authority, and populated by the consolidation of all the databases in various healthcare settings (national-level connectivity). In this regard, the POC devices need to be able to be automatically connected to a reliable and secure communication network or there must be a consistent and regular procedure to record the results and send them via qualified personnel.

Although Access to mobile phone technology has rapidly expanded in low and middle-income countries [26]–[29], most **POC** devices did not have the capability to connect to any communication network since connectivity was not considered as a priority at design time (or production time) of the devices. In this case, often the operator of the POC device needs to write down/copy-paste the results manually and then send them (directly or indirectly) to MOH (via mail or email or the internet, and so on). The manual procedure is error-prone and results in an increase in turnaround time (at local-level). This issue slows down analysis of the data at national-level as well since data at national-level are consolidated from all laboratories. This issue gets worse when portable POC devices are used. The closer POC diagnostic testing gets to the patient, the harder it becomes to consolidate data so that national-level authorities (like MOH) can analyze health outcomes countrywide [11], [30]. In addition, many laboratories are private and they just send data back to the hospitals and clinics. In other words, they do not send data to MOH. This is another issue at the national-level.

If reliable connectivity of POC devices is implemented, the time of transmission of test results from laboratories to clinics as well as national-level authority (like MOH) can be considerably reduced, human errors are eliminated, and a centralized database of all historical test results can be created (at both local-level and national-level) therefore decisions can be made without waiting for the data to be transmitted (figure 1).

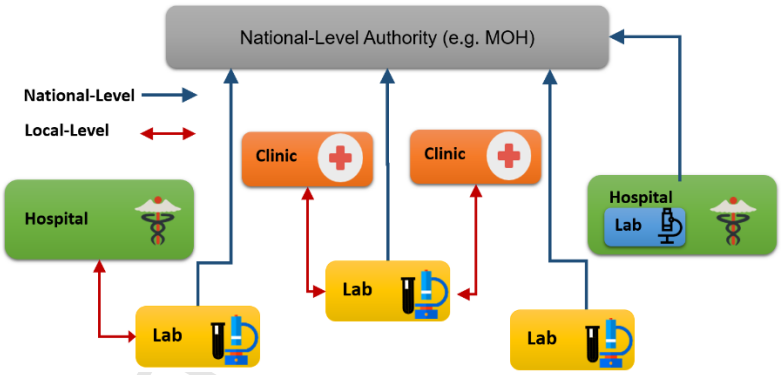


Figure 1: Ideal flow of test results at local-level and national-level

In the area of device connectivity, some companies have been working on the challenge of enhancing connectivity to encourage uptake of their diagnostics and to gain a larger share of the market. Recently data connectivity of POC devices has been changed from a “nice to have” feature to a “must have” feature in many cases. Some POC device manufacturers utilize built-in modems to send data. Some other companies use external modems to connect the devices to the mobile communication network. A few companies use the above-mentioned method to send data to a central database which is deployed in the cloud. Using cloud-based storage has many technical as well as cost advantages over traditional on-premises storage approaches. Providing data connectivity is usually the responsibility of device manufacturers. However, providing the communication infrastructure is the responsibility of customers. Technology push and demand pull suggests that all POC devices will have data connectivity features in the near future. However, the issue with communication infrastructure, especially in low and middle-income countries, persists.

The second issue, which is highly dependent on the first issue, is that of data analytics. The primary purpose of POC devices is to generate data about test results. The test result data items coupled with patient demographic data can be used to extract fine-grained surveillance information. In other words, the datasets composed of demographic data combined with test results can be used to get individual-level as well as population-level pictures of the health of patients. Such a picture forms the basis for the penetration and effectiveness of healthcare services and, therefore, efficiency of policies in the area of healthcare as well. In addition, by combining the above datasets with location data, mobility patterns (of humans, animals, and airflow that transmit diseases), trends in disease prevalence can be extracted and utilized for prediction, which is highly valuable for decision making.

In summary, availability of data (as a result of reliable connectivity, regulations, and policies) and data analytics are needed in order to support data-driven decision making which is the key procedure for monitoring and management of diseases and mitigation in case of diseases outbreaks.

In order to address the device connectivity requirement, some regulations need to: a) be proposed and introduced by national and international authorities, b) some technical specifications/standards need to be implemented by device manufacturers, and c) suitable network infrastructures need to be deployed by all healthcare settings. From the above requirements, each country is responsible for providing the network infrastructure. In low and middle-income countries providing the network infrastructure is an important problem. The proposed solution of this paper illustrates utilizing a variety of connectivity approaches to make IoT-enabled POC in order to resolve the connectivity issue when there is no local-level and/or national-level connectivity.

To address the data analytics challenge, cutting-edge big data technologies, which can manage and analyze the huge amount of data in batch (analysis on historical data) and real-time should be utilized. These technologies need an architecture with scalability, reliability, performance and fault tolerance characteristics. The implementation of the architecture using big data technologies provides the analytics infrastructure for large-scale management and analysis of data from multiple sources. In this case, POC devices can be considered as a data sources. However, access to data is a major challenge with a non-technical origin. The solution proposed by this paper shows how to provide the analytics infrastructure using state of the art big data architecture to get the most out of accessible data.

4 Data Generated by POC and Accessibility Issue

Combining the test results data (generated by POC) with patient demographic data results in comprehensive datasets. These comprehensive datasets can be used efficiently to extract fine-grained surveillance information about many diseases via data analytics at individual-level as well as at population-level. From a technical point of view in order to use all individual-level data analytics procedures, the datasets need to be available, usable and accessible. The comprehensive datasets are definitely valuable for extracting surveillance information, but they are only partially available and unfortunately are not accessible in most cases. The

comprehensive datasets in most cases can be collected from different sources. Most of the time data is stored in digital formats so it is possible to create a digital repository of the datasets and make them available. As it described in the previous section, eventually all POC devices will have data connectivity features and local-level and national-level connectivity will be in place in near future, even in low and middle-income countries if the network infrastructure is in place.

However, there are some serious issues in accessibility of the above-mentioned datasets. The comprehensive datasets are highly sensitive and have privacy-related issues. Because of myriad political, social, privacy and security issues, accessing individual-level data is very difficult (if not impossible). If demographic data are decoupled from test results, neither individual-level nor population-level information can be inferred from the test results. In other words, test results without patient demographic data lose their contexts and their use is limited to summary statistics (calculating count, average, minimum and maximum, for example) which has almost no value. Accessibility issue of comprehensive datasets is the most important barrier in front of extracting individual-level as well as population-level surveillance information. Unfortunately, the major reasons for accessibility issue are not technical and, in most countries, there is no feasible solution for this issue.

5 Proposed Solution

The POC devices are basically sensors and each sensor can generate large amounts of data during its intended measurement. In the case of POC devices, during the processing of a test sample, the devices generate lots of data which can be used to contextualize test results [31]. In this case, it is impossible to extract individual-level surveillance information. In other words, since the test results always contain identifiers (test identifiers), in theory, it is possible to join the test results with demographic data (using a patient identifier) to compile a comprehensive dataset even from POC deployment site. However, in practice, because of issues of accessibility of demographic data (due to its sensitive nature and potential patient identification risks), it is nearly impossible to link the test results to demographic data to generate a comprehensive dataset especially in low and middle-income countries. The lack of access to the comprehensive dataset is a major issue for extracting insights, especially at the individual level. POC generated data generally lack patient identifier data.

However, this lack of patient identification can be seen as an advantage at the population-level since it facilitates (or almost fully removes) getting ethical approvals at various hierarchical healthcare settings. In other words, while this is a major issue for extracting individual-level information, the POC generated data still can be used for extraction of useful population-level insights.

Since there is no need to get involved in working with highly sensitive data (patient demographic data), the population-level information can be extracted efficiently and quickly and, most of the time, without the need for approval of multiple ethics committees at different levels. This approach is a novel feature of the research described in this paper, and, to the best of the authors' knowledge, no other research project has been done with this feature. Using just data generated by POC devices (including test result and machine-generated data like, for

example, the duration of the test, the location of the device, errors and warnings and quality control parameters) it is possible to get population-level surveillance information without using sensitive data. In other words, the contribution of the solution of this research paper is a new type of knowledge extraction based on just POC machine generated data using big data analytics. A very important element of this type of knowledge extraction is that the POC machine generated data would never be identifiable, thereby protecting the rights and the anonymity of the individual, whilst still allowing for vital population-level evidence to be obtained.

In order to implement the proposed solution, data from various types of POC devices need to be managed and processed using a common data structure. Moreover, the solution needs to be able to run certain types of analytics and be extensible for future analytical needs. The common data structure and data analytics influence design and implementation of the system for the proposed solution. The following sub-sections explain the common data structure and analytics in more detail.

5.1 Common Data Structure of the Proposed Solution

The POC machine generated data are composed of a large set of data items about status and condition of the device during the test, and the result of the test, and the result of automatic quality control procedures during the test. Each type of device generates data in different structures, encoding, and formats. In this case, storage, management, and analysis of such POC machine generated data need a common data structure. The common data structure provides the unique model for mapping of different data structure and formats of various POC devices to a single semantic model.

For this research, a common data structure was designed for POC devices that record CD4 t-cell counts. In this case, a single unit of observation contains a hierarchy of data about the test (time, date, id, type of test, measurement of t-cell count), device (type, make, unique identifier and geographic location), quality control of the test (errors and warnings during test), consumables for the device (cartridge unique identifier) and operator of the device. The following figure shows a single unit of observation for a hypothetical test on a POC diagnostics device (for CD4 t-cell counts-HIV test).

```
{
    "test": {
        "id": "198033",
        "startTime": "12:00 01/05/2014",
        "endTime": "12:35 01/05/2014",
        "t-helper": "560",
        "QC": {
            "barcode": "passed",
            "expiry": "passed",
            "Volume": "passed",
            "reagent": "passed"
        },
        "Assay": {
            "id": "2",
            "assayInfo": " PIMA CD4 "
        },
        "Cartridge": { "-serialNo": "514231" },
        "Device": { "-serialNo": "PIMA-A-0037552" ,
          "position": {"lat":-23.6615, "long":22.7966}},
        "Operator": { "-id": "3198" },
        "software": { "-version": "2.1.9" },
        "Errors": {

        }
    }
}
```


Figure 2: Hierarchical data structure of a single unit of observation for t-cell count test (HIV-related immune system test)

5.2 Data Analytics in the Proposed Solution

In general, there are four categories of analytics: descriptive, diagnostics, predictive and prescriptive [32]–[34].

Descriptive analytics is used to explain what was/is happening in a given situation. This class of analytics can be used to answer questions such as how many tests are done each day, week, month or in real-time. How many tests has a certain POC device run in weekdays? How many types of errors occur? What are the most dominant error types? What are the trends in test results in certain POC device?

Diagnostic analytics helps in understanding why certain things happened and what are the key drivers. For example, a national health authority could use this type of analytics to answer questions such as why a certain type of POC error is increasing. Why a specific device is not working at its highest potential, or why all test results showing CD4 <500 are coming from a single device.

Predictive analytics helps to predict the future based on current and past situations (historical data). It is used to predict the probability of an uncertain outcome. For example, it can help to answer the following questions: what would be the growth/decay rate of the number of HIV-positive patients based on recent trends? How many tests are going to be done in next 2 months (for preventing cartridge stock out)? Where are additional resources (such as cartridges, devices, operators, etc.) needed for next 6 months?

The prescriptive analysis will suggest the best course of action to take to optimize the outcomes. Typically, prescriptive analysis combines a predictive model with domain specific rules. For example, it can suggest the best location for deploying a mobile POC devices based on existing POC devices, their capacities, population and spatial connectivity between POC devices (road network).

From analytics point of view, a system for analysis of the POC generated data requires specific considerations and solutions in order to implement all four types of analytics. In general, the system needs to be able to manage and analyze data in two different modes: batch mode and real-time mode. In this research, an architecture for management and analysis of POC generated data was designed based on best practices in big data system architecture and implemented via a combination of several big data technologies in a cloud environment.

6 Big Data Architecture of the Proposed Solution

Big data is defined as high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and processes optimization [34], [35]. This definition focuses not only on the characteristics of data but also on the way that the data are processed [36]. In fact, the scale of the data, flexibility in the integration of data from different sources and the distributed storage and processing of data in real-time are the main driving forces for the big data analytics technologies [36], [37].

Management and analysis of the POC generated data require specific considerations and solutions. In general, the system needs to be able to manage different forms and analyze a large volume of data in two different modes: batch mode and real-time mode.

Batch processing is used for getting a holistic view of a complex system which means the processing occurs after almost all the data items ingested into the system in a specified time. Batch processing is often complex and because of this it usually takes a long time to complete (few minutes to several days)[38]**. In contrast, in real-time mode, the processing and output generation of data is continuously done as soon as new data items are ingested into the system**[39]**. Commonly real-time processing uses a window of time (or time limit) for processing of the most recent ingested data in the system**[40]**. Usually the time window is a few minutes, or less. The main difference between batch and real-time processing is determined by the time of analysis of data. If the analysis of data is done long after the events occurs it is batch mode. In contrast, when the analysis of data occurs at almost the same time that events happen, it is a real-time mode** [41]**.**

In the context of this research, there are many situations that a full historical dataset of all POC devices is needed to be accessed in order to process data. Processing of full data always takes a long time. In this case, processing of data needs to be done in batch mode. For example, in order to perform statistical tests, predictive analytics and producing pivot reports, analysis of data need to be done in batch mode. **As another example training predictive analytics models and evaluation of their performance almost always needs batch processing (use of historical data).** In addition, for providing data to other systems like Geographical Information Systems (GIS) data need to be provided in batch mode.

On the other hand, time sensitive applications and sensor-based monitoring applications need real-time processing of data [42]**. For example, anomaly detection or event detection need a continuous evaluation of input data items against some criteria. Because of this, in the real-time processing, there is no need to access a large portion of historical data, but rather a small number of the most recent data inputs (table 1).**

Table 1: summary of purpose, data size, and type of analytics of batch and real-time processing

	Batch Processing	Real-time Processing
Main Purpose	**Processing/Analysing of all or large part of data**	**Processing/Analysing of most recent ingested data**
Size of Data	**Large batches of data**	**Individual records of data or micro-batches (a few records)**
Latency	**High (minutes to hours)**	**Low (milliseconds to few minutes)**
Analytics	**Complex Analytics (simulations and modeling, advanced machine learning, predictive analytics, statistical learning, etc.)**	**Simple and Complex Analytics (simple statistics like aggregate functions, descriptive analytics, anomaly detection, etc.).**

Designing systems capable of handling both batch and real-time processing is a complex task and requires an effective conceptual architecture for implementing the system. As an example, to achieve real-time processing capability a system must be able to perform message processing without having a costly (computationally) storage operation in the critical processing path. In this case, the system must have the capability to distribute processing across multiple machines to achieve scalability [42]**. Fortunately, with help of several big data technologies, it is possible to design and implement a system architecture that can handle batch as well as real-time processing in a unified way. One such conceptual architecture is lambda architecture.**

Figure 3 shows an architecture based on lambda architecture[43], [44] **that has been customized and extended for the purpose of this research. Lambda architecture is a conceptual, generic, scalable, and fault-tolerant data processing architecture based on distributed systems** [36]**. Lambda architecture is designed to handle massive quantities of data by taking advantage of both batch and stream processing** [45]**.**

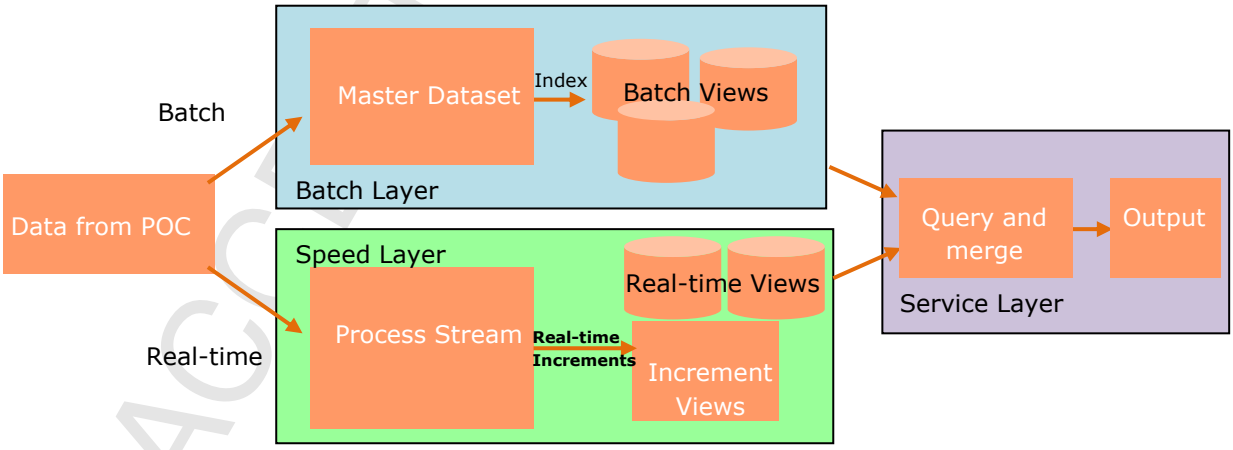


Figure 3: Big Data Architecture of the Proposed Solution

As is illustrated in Figure 3, in the above architecture data from various types of POC devices ingested into the system in two paths for two types of processing. There are three

layers in the architecture; batch, speed, and service. The batch and speed layers are responsible for performing batch processing and real-time processing respectively, and service layer provides access to data in either/both paths.

The batch layer is composed of a distributed append-only data storage system that stores data as time-stamped data items (master dataset in Figure 3). The master dataset intentionally does not support update or delete operations, meaning that all the data in the master dataset have a time identifier and new data do not override the existing data.

Based on most common types of analytics, the batch layer pre-computes batch views on the master dataset regularly in specified times. These batch views provide a high latency representation of whole data. Computations for building batch views usually are written like single-threaded programs, and because of this can be automatically parallelized across a cluster of machines. This implicit parallelization makes batch layer computations scale to datasets of virtually any size.

Same data ingested into the batch layer is also inserted temporarily in speed layer which is responsible for the real-time processing of data. Then, after a specific window of time, new data items replace the previously processed data which means only the most recent data is stored in the speed layer. The speed layer generates real-time views using real-time streaming data. In contrast to batch layer, the speed layer provides a real-time view of the ingested data. The major difference between batch and speed layer is that, in order to achieve the lowest possible latencies, the speed layer does not look at all the new data at once. Instead, it updates the real-time view as it receives new data instead of re-computing them like the batch layer does.

The service layer has the ability to query (access) either historical and real-time data for performing analytics, or to provide data to external analytical software. More importantly, when access to all data including historical and real-time data are needed, the service layer can merge data from batch and speed layers and provide the most complete and most recent data for analytics purposes.

For implementing the designed architecture, each POC device needs to send full generated data to a cloud infrastructure (Microsoft Azure) as messages. As mentioned before, many types of POC devices are not designed with this capability. Therefore, in order to resolve the connectivity issue, we used a few different methods.

In the first method, we developed a software capable of reading low-level details from the devices. Then the software could send the data messages to the cloud platform using USB modems and pre-paid sim cards with the help of regional mobile communication network (as SMS messages). We utilized this method for 65% of the devices. For remaining 35% devices, we used network port of the POC devices (second method). **The network port is usually utilized for sending the results of tests to a central computer (hub) for printing or sending the reports to the clinics or hospitals. In this research, we implemented a rather complex software background service called Resilient Queue System (RQS) to send the data over Internet to the cloud. The reason for the design and implementation of the RQS was**

unreliability of Internet in almost 35% of devices. In other words, during the time that the devices work, the Internet was available for only a few hours (partially connected situation). Therefore, it was not possible to use the first method and send the test results as the tests happen. We implemtened the RQS to ensure that all the data were received successfully in the cloud. In the RQS, results of tests are stored temporally as messages in a resilient queue, until the Internet connection becomes available. The system then sends the messages one-by-one (in a first in, first out manner). Sending messages does not stop until the network disconnects from the Internet.

On the cloud side, we implemented a service to receive messages and send an acknowledgment code to the RQS. As soon as the RQS gets the acknowledgment code, it removes (deletes) the message from the queue.

The above architecture was implemented using HDInsight family of technologies including Apache Strom, Hive, and HBase inside Microsoft Azure cloud computing platform (Figure 4). The HDInsight is Microsoft's managed implementation of Hadoop ecosystem [46]**. The HDInsight can run on server versions of Linux and Windows. In this research, the system utilizes Windows Server at the Operating System level.**

Data ingestion is done in the system using an ASP.NET Web API layer. Then for the batch layer, HDInsight storage service was used to store the data (using Azure blob storage) and Apache Hive was utilized to create and update batch views. Apache Storm was deployed for generating real-time views in the service layer and Apache HBase was used in the service layer. Finally, for end users, a dashboard for different batch and real-time metrics was created using the PowerBI. **Also for other software that solicit data from the system, APIs were provided using ASP.NET Web API.**

In addition to the dashboard, all data in the system is accessible to external analytical clients like R (for performing statistical analysis), Python (for performing machine learning) and ArcGIS (for performing spatial analysis). In section 10, the dashboards, results of a machine learning model for anomaly detection, and a spatial analysis for finding optimal location of a new mobile lab are illustrated.

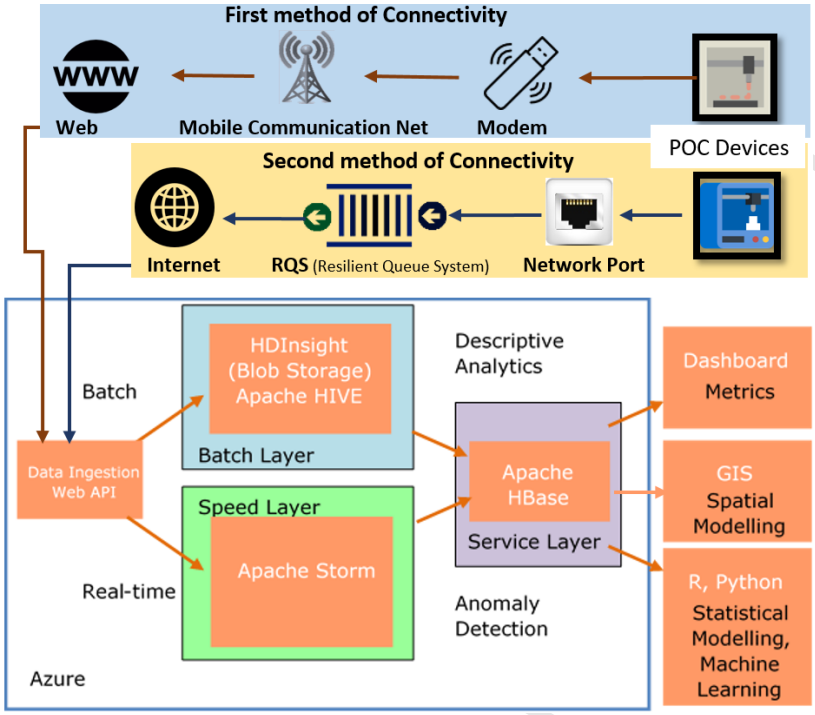


Figure 4: Implementation of proposed solution

7 Benefits of the method for research communities, health authorities, and device manufacturers

The proposed method provides the ability to perform all types of analytics (descriptive, diagnostics, predictive and prescriptive) using just POC generated data (unidentifiable data) to extract useful information and insights potentially for three groups; researchers, healthcare settings, and device manufacturers.

To the best of authors' knowledge the current paper is the first research paper about using just POC machine generated data for scientific purposes. Identification of trends in the distribution of diseases, evaluation of influential factors in disease spread, pinpoint the high-risk areas and population (for fine-grained confirmatory analysis) are a few examples of opportunities for research communities.

For device manufacturers, POC machine generated data sometimes have been used for monitoring the operational status and performance of devices (from quality control perspectives). However, POC generated data hardly ever have been utilized for other useful applications such as modernizing supply chain management, planning the required equipment and consumables, and predictive (or preventive) maintenance of devices.

For healthcare settings (labs, clinics, hospitals as well as national authorities) the method can be utilized for improving internal workflow (such as monitoring the performance of lab facilities and device operators), eliminating latency in availability and visibility of test results, providing quality control visibility of all settings, preventing stock-outs (of cartridges for example), optimizing usage of the devices, dynamic resource allocation

(finding best locations for deploying mobile labs), and enabling further health-related analysis of data.

From the implementation point of view, the system has been designed in a way that can handle a huge amount of data from different data sources in both batch and real-time modes. Since the system is deployed to the cloud, there is no need for the establishment of highly expensive data centers or buying expensive servers. The fact that the cloud provider company (Microsoft inc. in the case of this research) is responsible for providing security, availability, recovery, and maintenance of the system is also a benefit. The horizontal scalability is done automatically by the cloud provider (this is called an elastic computing cloud). The horizontal scalability is one of the advantages of the modern architecture of the implementation of the method proposed in this research over conventional healthcare data management and analysis systems. In general, most conventional data analysis techniques are not suitable for working with huge amounts of data from different sources and with various data structures. However, as illustrated in the paper, careful design and implementation of the system enable benefiting from all advantages of the cloud.

In addition, the implemented system can deliver data to external analytical software for further analysis. For example, it is very easy to connect to the system and get data in environments such as R and Python (for statistical learning and machine learning), ArcGIS (for spatial analysis and geostatistics), and Tableau (for visual analytics). The following section provides a broad range of analytics using the POC generated data for three above mentioned categories.

8 The implemented Data Analytics and Dashboards

For testing the functionality of the implemented system data from more than 2000 POC, devices in 3 countries in Africa ingested in the system (all from 2012-2014). The **Data were gathered from national health authorities in those countries, African Health Observatory (AHO), The Global Health Network, device manufacturers and private companies (a few private consultancy companies like, for example, Alere and Cepheid), and WHO.** Also, a large amount of simulated data were continuously generated and ingested in the system to test the real-time and batch processing capabilities and analytics of the system. Several types of data analytics were implemented.

Table 2, shows details of data input, and data analysis in the implemented system. Since data records of the POC machine data are stored and analyzed as events (time-series observations), the table contains two categories of analytics that cover the full spectrum of analytics; temporal and spatio-temporal. The main difference between the two types of analysis is that in spatio-temporal data analysis, location (position) and spatial relationships (like distance, direction, and connectivity between locations of the devices) were used for grouping and clustering the data whereas in temporal data analysis time of the test is mostly used for that purpose.

Table 2: Implemented Analytics on POC generated machine data

Type	Data Input	Analytics
Temporal	Historical and Real-Time information for all sites, group of sites and individual sites	total successful tests, total errors, average tests per month (week, day and hour), average successful tests, tests statistics, error rate, top error types, the total number of operators, relationships among errors and number of tests through different types of regressions
	Health-Related Measurements and Metrics	Percentage of patients with CD4 <200 (some important values for CD4 are 500, 350 and 200), real-time monitoring of sites with average CD4<500 or monitoring sites which has large number of patients with CD4 < 350, Number of tests with CD4 <200 in order to estimate the needed resources for ART (Antiretroviral Therapy), **providing personalized advice after tests (if a professional was available and results were ready).**
	Stock Information	Report of daily consumption of resources (for each device or site), current stock value (for each site), total stock used, estimated daily consumption, estimated stock out date for each site, **the prediction of consumption of cartridges for next two months**, efficient **allocation of cartridges (transporting cartridges from sites with low demand to sites with high demand),** warning and alert long before stock out.
	Operator-related Information	Real-time performance monitoring of operators, determining most and least precise operators **for retraining purposes** (with errors that are related to the operator like expiry date of cartridges).
Spatio-Temporal	Location, population, mobility, connectivity of devices (by transportation networks)	Spatial distribution of sites, service area determination of sites, **and optimized resource allocation** for finding the best place to deploy mobile clinics (based on location, distance, population and existing transportation network).

The result of above analytics in the implemented system is a set of interactive dashboards. **There are three main dashboards (Healthcare Settings, Test Analytics, Errors and Warnings) in the system. Each dashboard consists of a set of interactive panels of visualizations and reports.** The **following illustrations display separate visualizations and reports of some of the dashboards, and describe how some useful information can be extracted based on the POC generated data. Figures 5 and 6, illustrate parts of "Test Analytics" and "Healthcare Settings" dashboards respectively.**

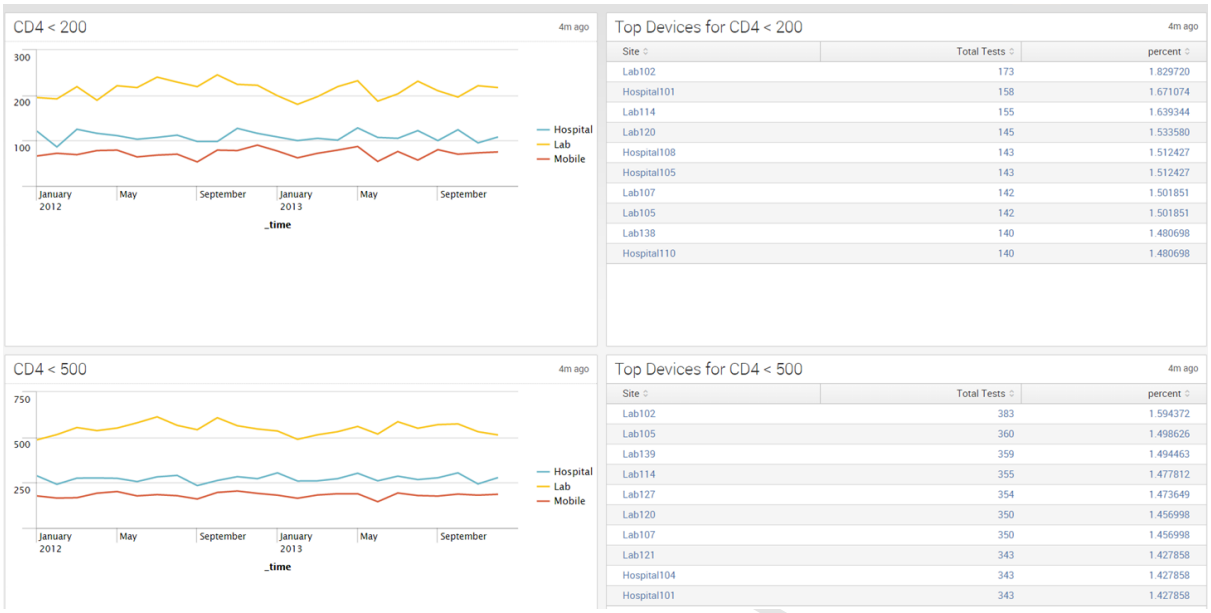


Figure 5: Part of Test Analytics dashboard. Two left panels show the count of tests with CD4 < 200 and CD4 < 500. On right panels, devices with highest number of results with CD4 < 200 and CD4 < 500 are shown (each device has a location).

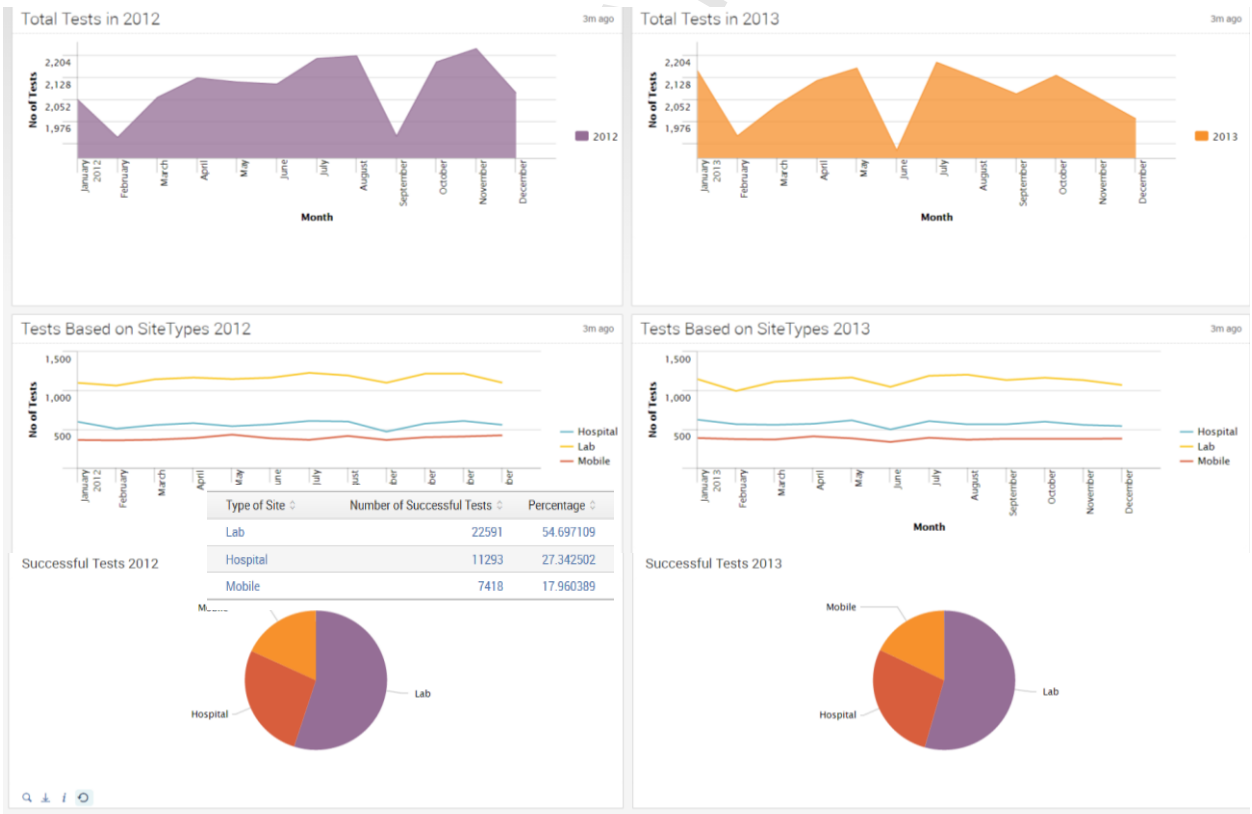
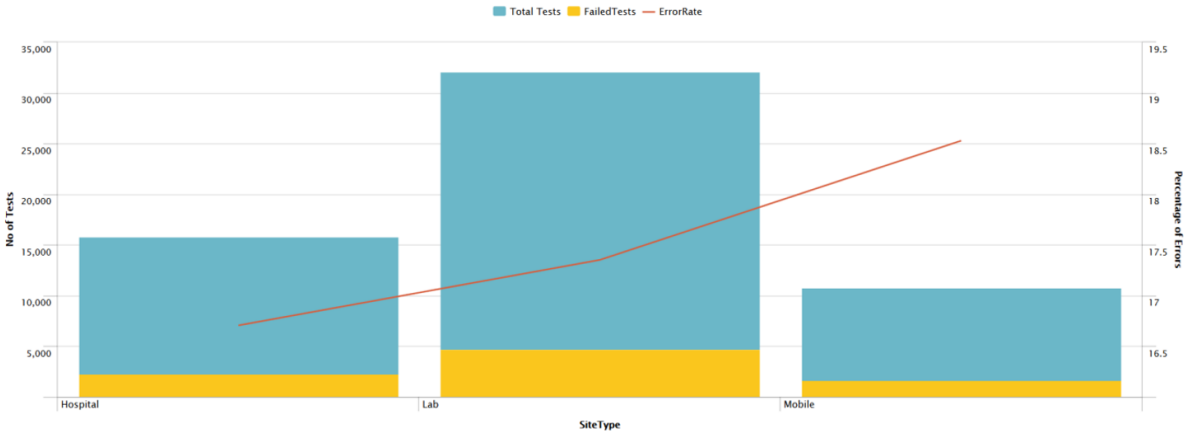


Figure 6: Part of Healthcare Setting dashboard. The two left panels illustarate the count of tests and successful tests in different healthcare settings. Hovering mouse over the pie chart shows the count and percentage of the tests grouped by healthcare setting type.

Figure 6 show some information about successful and failed tests in different healthcare settings. More than half of tests (54.7%) had been done in labs and about 27.3% and 18% were performed at hospitals and mobile labs respectively. Based on Figure 7, in spite of the fact that a number of tests in labs are almost three times more than the number of tests in mobile labs, the percentage of failed tests in mobile labs is more than labs and hospitals as well. This interesting observation can be seen as the red line in Figure 7. This line can be interpreted as a result of one or combinations of following reasons: 1) Non-uniform allocation of human resources e.g. relatively inexperienced operators in mobile labs, 2) Allocation of faulty devices to mobile labs (or any problem associated with transportation of devices) and 3) Existence of a serious maintenance issue due to environment; for example, maintaining cartridges in environments with more than 40 degree centigrade for devices or test sample in mobile lab settings.



SiteType	Total Tests	FailedTests	ErrorRate
Hospital	13558	2265	16.706004
Lab	27334	4743	17.352016
Mobile	9105	1687	18.528281

Figure 7: Total Tests (blue bars), Failed Tests (yellow bars) and Failed Tests Percentage (red line) for each Site Type

Figure 8, illustrates the monthly trend of tests (successful tests per months for all sites) in 2014. January to February, August to September and November to December show decreases. Similarly, February to March, June to July and September to October show increases in tests. These trends might have occurred randomly, or because of cultural, social or environmental factors. Further studies need to be done for any causal statements about this observation.

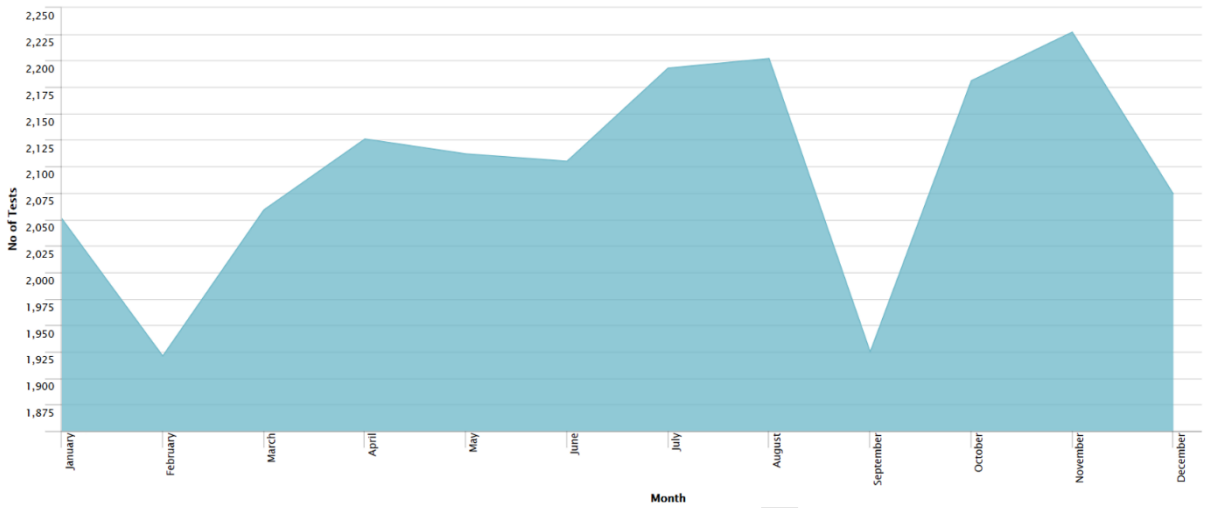


Figure 8: Trend of Number of tests per months for 2014

Figure 9 represents the monthly trends of tests based on site type. As illustrated in the figure, unlike the general monthly trend of Figure 8, in mobile labs the number of tests increased during April to May, and the November to December periods.

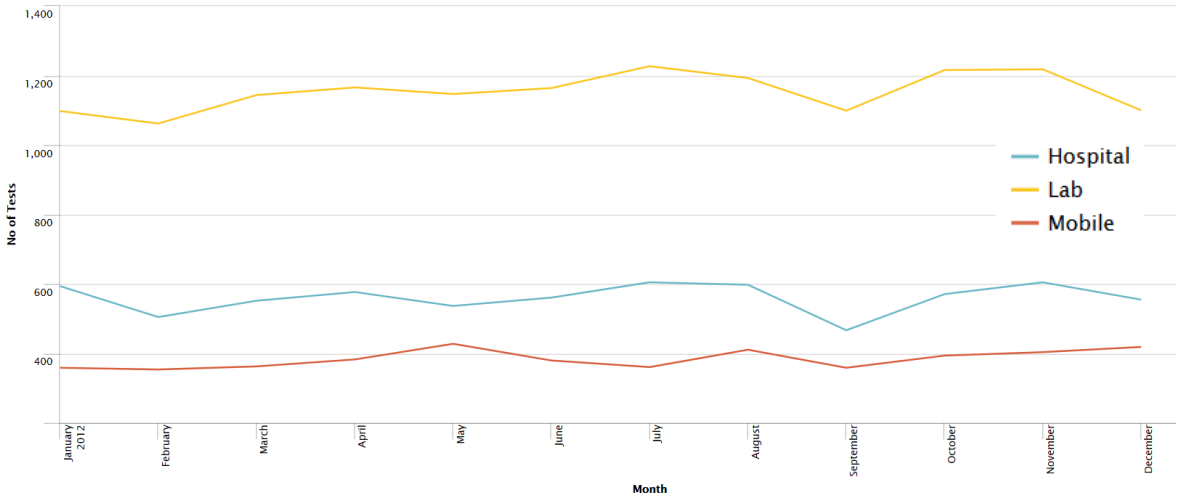


Figure 9: Trend of Number of tests per months for 2012 for each Site Type

The monthly trends (Figure 8, 9) for different years, can be compared and general anomalies can be detected. Further research and observations are needed in order to determine reasons for the causality of the anomalies. However, this can be seen as an advantage of the method proposed in this paper, especially in real-time data collection. It is possible to extract norms of the data in different seasons based on historical data and then using some rules for anomaly detection or using artificial intelligence techniques for the system to learn from data. The system then is capable of detecting anomalies automatically without the need for human interaction. This anomaly detection can be done in real-time. In this case, the system can take appropriate actions by using alerts or triggers, for example, by broadcasting an email to responsible people to make them aware of the current situation. In other words, the value

remains in that these anomalies could be determined in real time and not require the delay that validation of traditional surveillance methods usually entails. In addition, these anomalies can result in automatic actions (like an alert, or in more serious situations shutting down of the device).

Figure 10 shows the pie chart for errors. As is illustrated in the figure, human error (errors related to the manual tasks of operators) accounts for 77% percent of all errors. In other words, using expired cartridges (cartridge expiry) and putting the inadequate amount of blood in the cartridge (volume errors) account for 77% of all errors. This is a highly valuable observation since it means operators may need more training about using and maintaining the devices and cartridges in appropriate conditions. At the other hand, this important observation means with relevant actions (like more training) it is possible to eliminate about 77% of total errors. Again, the effectiveness of relevant actions can be monitored in real-time, with the impact on the quality of tests being observed immediately.

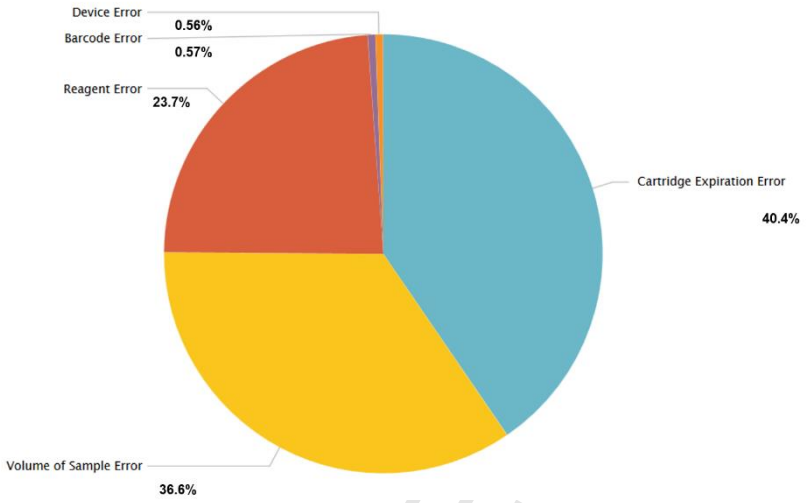


Figure 10: Major Error Types

As it stated above, the mentioned valuable observation about error types might be because of some maintenance issues. For example, the POC devices and their cartridges have to be maintained and used in a certain range of temperature (2-30 centigrade) or humidity (10%-75%). With further investigations about weather condition from multiple sources, just 12% of measurements were done outside of the temperature and humidity range. Therefore, oversimplifying of the maintenance and usage procedures of POC devices might be the cause of this error. This might lead us to another interesting fact. International health organizations like CDC and WHO tend to oversimplify the usage and maintenance procedures of various POC devices in resource-limited settings. This can be seen in the following quote “The characteristic of the Pima CD4 assay indicates that it is ideal for a point-of-care or resource-limited setting. The assay requires minimal training and technical skill with no sample processing manages results and reports on the analyzer, and does not require maintenance, cold-chain storage, or ancillary equipment.”[47]. Although enthusiasm and hope are increasing around POC diagnostics for diseases of global health importance, a deeper appreciation of

likely barriers in each healthcare setting and operations (for example lack of training) might help test developers and public health managers to identify the complexities of using POC [48].

Figure 11 shows the monthly trends of errors. This reveals a remarkable observation; the errors seem to be based on repeating patterns in every successive two or three months. This might be due to the pattern of the inspection procedure every successive two or three months. As another observation, before the new years' holidays, error rates increase while after holidays error rates decrease. This can be hypothesized as most of the experienced operators tend to go to new years' holidays before the New Year eve. These kinds of patterns can be identified and detected by defining rules or inferring patterns in the system. Using the mentioned rules or patterns, the system can identify similar trends automatically in real-time and then inform the responsible people (via email for example) in order to take appropriate actions.

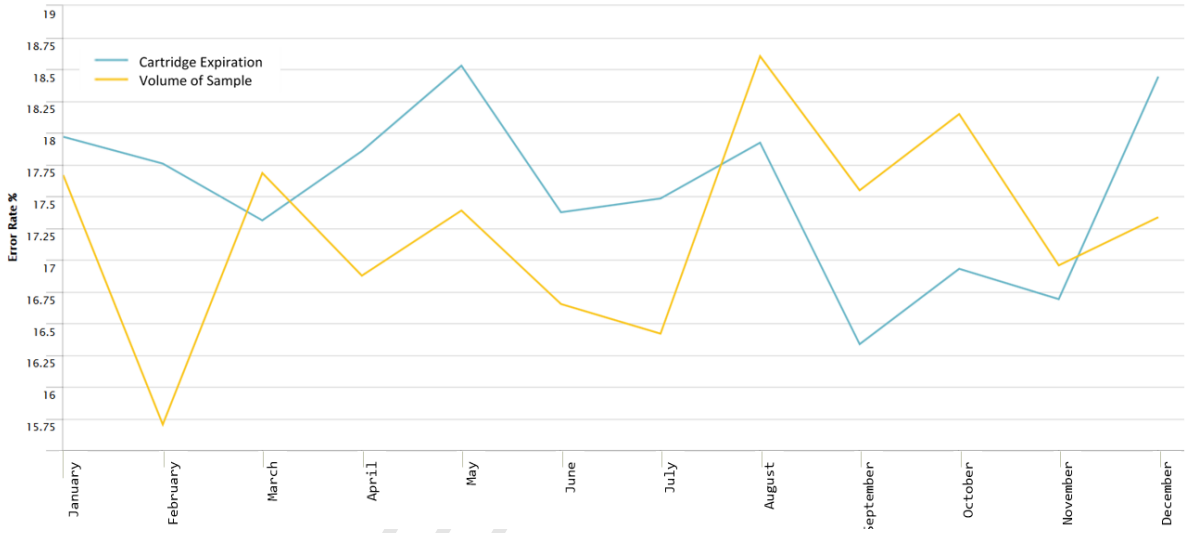


Figure 11: Error Rate per months

Treatment guidelines generally recommend that HIV-infected patients with a CD4 cell count less than 350 should receive highly active antiretroviral therapy (HAART) [16]. Figure 12 shows that mobile labs have the most of the tests with CD4 less than 350. This information can be used for identification of population or areas at high risk, or to allocate more resources (in this case enough HAART therapy or HAART therapy adherence support) to those sites.

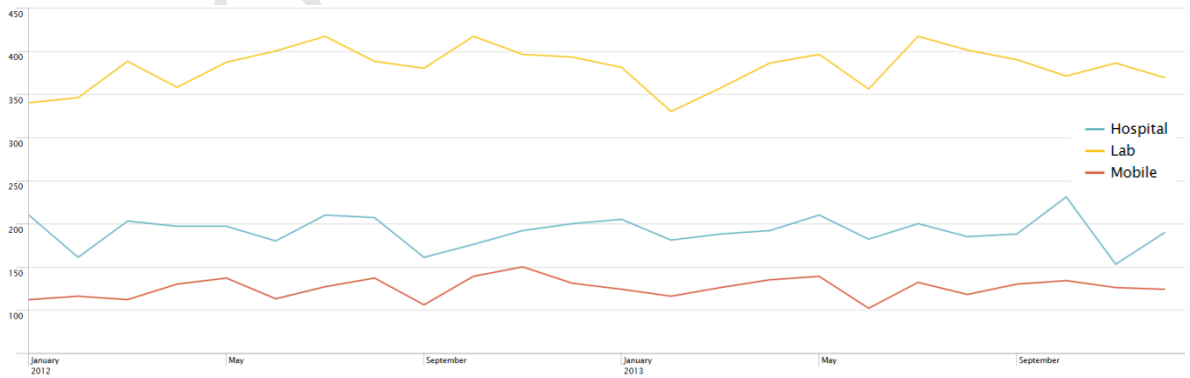


Figure 12: CD4 < 350 grouped by site type per month

It is possible also to create similar reports for other amounts of CD4 for different purposes. It is also easy to calculate simple statistics for any numeric or categorical variable in the dataset. Figure 12, shows the descriptive statistics for CD4 variable. As it is shown, the amount of CD4 in 25% of all tests are less than 218 (First Quartile).

Minimum	12
First Quartile	218
Median	432
Mean	475
Third Quartile	644
Maximum	2102

Figure 13: Summary Statistics for CD4 (25% of test results are below 218, 25% are more than 644)

As it described before, these reports can be used effectively to identify high-risk areas and population. Figure 14 illustrates the top 10 devices (and therefore sites) with the highest number of CD4 less than 500. Finding the location of the devices is an easy task since the location is part of stored data in the common data structure.

Site	Total Tests
Lab102	383
Lab105	360
Lab139	359
Lab114	355
Lab127	354
Lab120	350
Lab107	350
Lab121	343
Hospital104	343
Hospital101	343

Figure 14: Devices with the highest number of CD4 <500

Figure 15 shows operator ID codes of those who made lowest mistakes (defined as failed tests due to human error). The similar report could show the operator ID and related sites which made the highest number of failed tests. The latter report can be used for providing more training or frequent inspection (or unannounced inspections).

OperatorID	count
157	164
323	162
96	160
362	157
93	156
40	156
311	156
231	156
39	155
187	154

Figure 15: Top 10 Most Precise Operators

Figure 16 shows ten devices with the highest number of tests. This report can be used for regular checks and overhaul maintenance planning of the devices. Also, the similar report can show which devices are mostly idle. The advanced type of analytics (optimization) can use the mentioned report for producing a plan for allocation of idle devices or consumables to the sites which need more capacity.

Device	Total Tests
Lab121	621
Lab127	597
Lab108	596
Lab120	595
Lab139	594
Hospital112	594
Lab107	592
Lab102	592
Lab105	591
Hospital111	591

Figure 16: Top 10 Most Used Devices (regular checks is more needed than other devices)

The location of devices is an important data item in the common data structure. The locations of devices can mostly be considered as the location of hospitals and labs for which the tests are done. This assumption can be automatically controlled with the IP (Internet Protocol) address of the network (via geolocation API of HTML5). For mobile labs (and the devices with sim modems), there are several techniques for determining the location of the devices. Generally using cell towers location in a communication network and techniques like triangulation are precise enough to calculate the location of mobile labs automatically. A good example of prescriptive analytics is to find the most suitable location for a new mobile lab. For performing the site suitability analysis, population data of the country, road network, and location and capacity of existing labs were utilized as input to a complex spatial analysis model for determining a most suitable location in a GIS environment (figure 17).

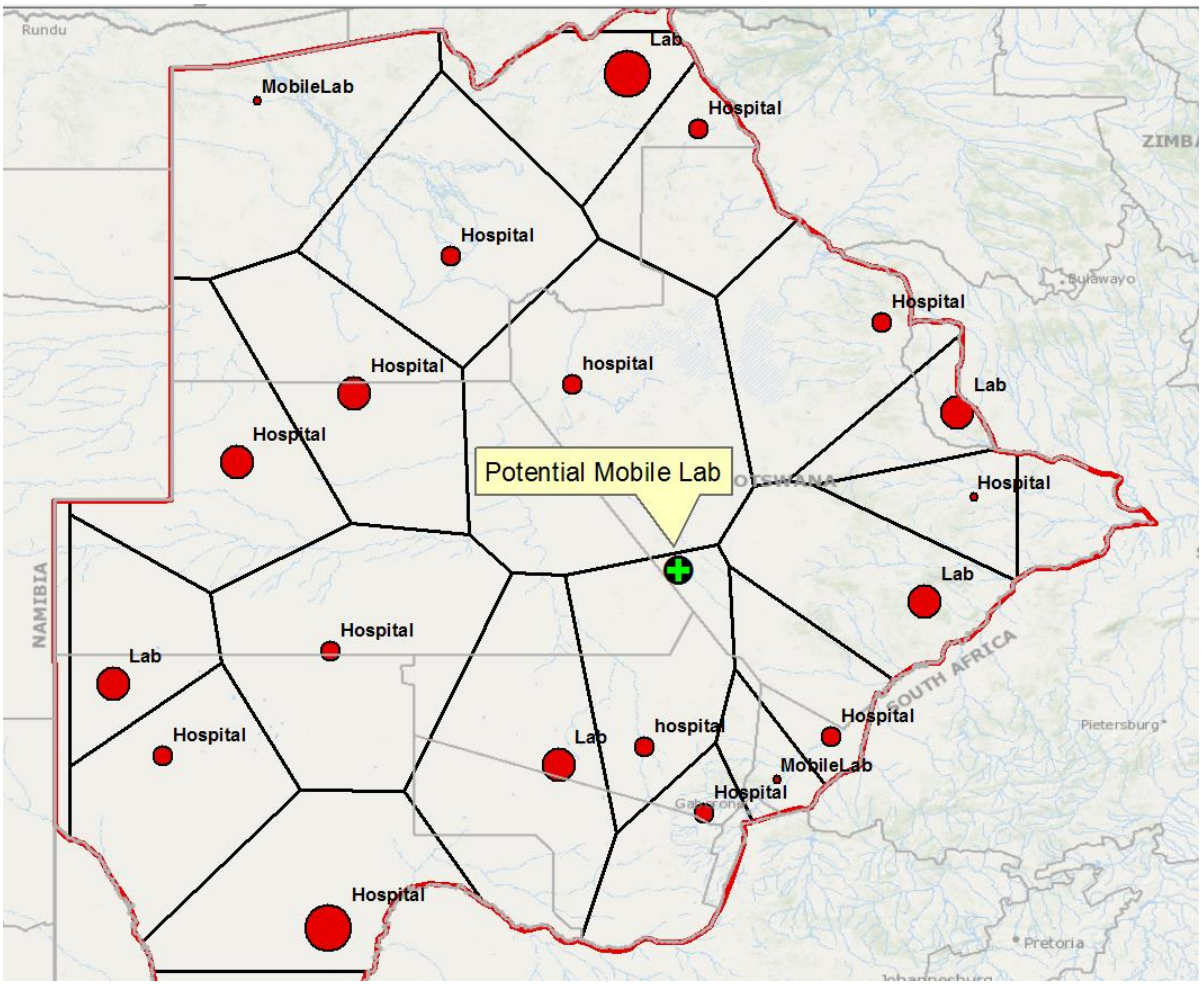


Figure 16: Most suitable location for a new mobile lab
(size of red symbols are proportional to capacity of sites)

9 Conclusions and Future Work

This research paper describes a novel approach for knowledge extraction for healthcare domain. Usually demographic data about each patient is needed for performing complex analytics. Unfortunately, because of a mixture of political, social, privacy, technical, and security issues, especially in low and middle-income countries, using demographic data at the individual level is very difficult (if not impossible). The contribution of the paper is the performance advanced analytics to extract valuable insights without using demographic data at the individual level. An important and unique aspect of this novel approach is the data are not individually identifiable so there is no concern about privacy. In this research, just device (POC) generated data has been used to extract valuable and useful information and insights at a population-level. The proposed method provides the ability to perform all types of analytics (descriptive, diagnostics, predictive and prescriptive) using just POC generated data (unidentifiable data) to extract useful information and insights potentially for researchers, healthcare settings, national and international health authorities, and device manufacturers. Identification of high-risk

areas and high-risk populations are two important results of applying the method described in the paper.

Also, the paper describes the implementation of the method, using a large-scale and complex architecture using IoT-enabled POC devices and big data analytics. As illustrated in the paper, the implemented system can perform descriptive, diagnostic, predictive, and prescriptive analytics in batch and real-time modes.

Combining POC machine data with other sources of data, especially volunteer content generation applications such as social microblogging (like Twitter), social networking (like Facebook), location-based services (like Foursquare), and volunteer geographic information (like OpenStreetMap) is a future direction for this research. With **these data sources and by using cutting-edge methods from artificial intelligence and machine learning, a new set of analysis can be done on the data and therefore new insights can be extracted from different types of analytics.**

10 References

[1] P. K. Drain *et al.*, "Diagnostic point-of-care tests in resource-limited settings," *Lancet Infect. Dis.*, vol. 14, no. 3, pp. 239–249, 2014.

[2] R. Peeling, "Bringing diagnostics to developing countries: an interview with Rosanna Peeling," *Expert Rev. Mol. Diagn.*, vol. 15, no. 9, pp. 1107–1110, Sep. 2015.

[3] M. Urdea *et al.*, "Requirements for high impact diagnostics in the developing world.," *Nature*, pp. 73–9, Nov. 2006.

[4] C. C. Y. Poon, Yuan-Ting Zhang, and Shu-Di Bao, "A novel biometrics method to secure wireless body area sensor networks for telemedicine and m-health," *IEEE Commun. Mag.*, vol. 44, no. 4, pp. 73–81, Apr. 2006.

[5] M. Fiordelli, N. Diviani, and P. Schulz, "Mapping mHealth research: a decade of evolution," *J. Med. Internet*, 2013.

[6] M. Kay, J. Santos, and M. Takane, "mHealth: New horizons for health through mobile technologies," *World Heal. Organ.*, 2011.

[7] T. A. Okoror, R. BeLue, N. Zungu, A. M. Adam, and C. O. Airhihenbuwa, "HIV Positive Women's Perceptions of Stigma in Health Care Settings in Western Cape, South Africa," *Health Care Women Int.*, vol. 35, no. 1, pp. 27–49, Jan. 2014.

[8] C. A. Heimer, "'Wicked' ethics: Compliance work and the practice of ethics in HIV research," *Soc. Sci. Med.*, vol. 98, pp. 371–378, 2013.

[9] J. Sugarman, S. M. Rose, and D. Metzger, "Ethical issues in HIV prevention research with people who inject drugs.," *Clin. Trials*, vol. 11, no. 2, pp. 239–45, Apr. 2014.

[10] J. M. Kirigia, A. Seddoh, D. Gatwiri, L. H. Muthuri, and J. Seddoh, "E-health: determinants, opportunities, challenges and the way forward for countries in the WHO African Region," *BMC Public Health*, vol. 5, 2005.

[11] L. Oldach, A. Sall, J. Lehe, and P. Fernandes, "ASLM Challenges & Implications for POC

Diagnostics in Africa," *ASLM*, no. 13, 2015.

[12] S. D. Lawn *et al.*, "Advances in tuberculosis diagnostics: the Xpert MTB/RIF assay and future prospects for a point-of-care test," *Lancet Infect. Dis.*, vol. 13, no. 4, pp. 349–361, 2013.

[13] G. W. T. Michael F. Drummond, Mark J. Sculpher, Karl Claxton, Greg L. Stoddart, *Methods for the Economic Evaluation of Health Care Programmes*. Oxford University Press, 2015.

[14] L. J. Abu-Raddad, P. Patnaik, and J. G. Kublin, "Dual infection with HIV and malaria fuels the spread of both diseases in sub-Saharan Africa," *Science (80-.).*, vol. 314, no. 5805, pp. 1603–1606, 2006.

[15] S. D. Lawn, A. D. Kerkhoff, M. Vogt, and R. Wood, "Diagnostic accuracy of a low-cost, urine antigen, point-of-care screening assay for HIV-associated pulmonary tuberculosis before antiretroviral therapy: a descriptive study," *Lancet Infect. Dis.*, vol. 12, no. 3, pp. 201–209, 2012.

[16] W. H. O. E. R. Team, "Ebola Virus Disease in West Africa — The First 9 Months of the Epidemic and Forward Projections," *N. Engl. J. Med.*, vol. 371, no. 16, pp. 1481–1495, 2014.

[17] W. H. Organization and others, "Urgently needed: Rapid, sensitive, safe and simple Ebola diagnostic tests." Geneva, Switzerland. Retrieved from http://www. who. int/mediacentre/news/ebola/18-november-2014-diagnostics/en, 2014.

[18] A. Sanchez *et al.*, "Analysis of human peripheral blood samples from fatal and nonfatal cases of Ebola (Sudan) hemorrhagic fever: cellular responses, virus load, and nitric oxide levels," *J. Virol.*, vol. 78, no. 19, pp. 10370–10377, 2004.

[19] N. F. Walker *et al.*, "Evaluation of a point-of-care blood test for identification of Ebola virus disease at Ebola holding units, Western Area, Sierra Leone, January to February 2015," *Ebola virus Dis.*, p. 64, 2015.

[20] W. Stevens, N. Gous, N. Ford, and L. E. Scott, "Feasibility of HIV point-of-care tests for resource-limited settings: challenges and solutions," *BMC Med.*, vol. 12, no. 1, pp. 1–8, 2014.

[21] K. Lewandrowski, "Point-of-care testing: an overview and a look to the future (circa 2009, United States)," *Clin Lab Med*, vol. 29, 2009.

[22] I. V Jani and T. Peter, "How point-of-care testing could drive innovation in global health," *New Engl J Med*, vol. 368, 2013.

[23] R. W. Peeling, "Diagnostics in a digital age: an opportunity to strengthen health systems and improve health outcomes," *Int. Health*, vol. 7, no. 6, pp. 384–389, 2015.

[24] World Health Organization, *Tuberculosis Diagnostics Technology and Market Landscape 3RD EDITION*. 2014.

[25] D. McNeill and T. H. Davenport, *Analytics in Healthcare and the Life Sciences: Strategies, Implementation Methods, and Best Practices*. Pearson Education, 2013.

[26] C. B. Aranda-Jan, N. Mohutsiwa-Dibe, and S. Loukanova, "Systematic review on what works, what does not work and why of implementation of mobile health (mHealth) projects in Africa," *BMC Public Health*, vol. 14, no. 1, p. 188, 2014.

[27] N. Leon, H. Schneider, and E. Daviaud, "Applying a framework for assessing the health system challenges to scaling up mHealth in South Africa," *BMC Med Inf. Decis Mak*, vol. 12, 2012.

[28] M. Blackenberg, C. Worst, and C. Scheffer, “Development of a Mobile Phone Based Ophthalmoscope for Telemedicine.” 2011.

[29] K. De Tolly, D. Skinner, V. Nembaware, and P. Benjamin, “Investigation into the use of short message services to expand uptake of human immunodeficiency virus testing, and whether content and dosage have impact,” *Telemed J E Heal.*, vol. 18, 2012.

[30] R. W. Peeling and R. McNerney, “Emerging technologies in point-of-care molecular diagnostics for resource-limited settings,” *Expert Rev Mol Diagn*, vol. 14, 2014.

[31] T. O’Reilly, J. Steele, M. Loukides, and C. Hill, “How Data Science Is Transforming Health Care Solving the Wanamaker Dilemma,” pp. 1–29, 2012.

[32] G.-Y. Vahn, “Business analytics in the age of Big Data,” *Bus. Strateg. Rev.*, vol. 25, no. 3, pp. 8–9, 2014.

[33] B. Baesens, *Analytics in a big data world: The essential guide to data science and its applications*. John Wiley & Sons, 2014.

[34] J. Dean, *Big Data, Data Mining, and Machine Learning: Value Creation for Business Leaders and Practitioners*. Wiley, 2014.

[35] P. Amirian, A. Basiri, F. Van Loggerenberg, T. Moore, T. Lang, and M. Varga, “Intersection of Geospatial Big Data, Geocomputation and Cloud Computing,” in *1st ICA European Symposium on Cartography*, 2015, pp. 72–74.

[36] P. Amirian, F. Van Loggerenberg, T. Lang, and M. Varga, “Geospatial Big Data for Finding Useful Insights from Machine Data,” in *GISResearch UK 2015*, 2015.

[37] J.-G. Lee and M. Kang, “Geospatial Big Data: Challenges and Opportunities,” *Big Data Res.*, vol. 2, no. 2, pp. 74–81, 2015.

[38] P. Amirian, A. Basiri, and A. Winstanley, “Evaluation of Data Management Systems for Geospatial Big Data,” in *Computational Science and Its Applications – ICCSA 2014*, vol. 8583, B. Murgante, S. Misra, A. M. A. C. Rocha, C. Torre, J. G. Rocha, M. I. Falcão, D. Taniar, B. O. Apduhan, and O. Gervasi, Eds. Springer International Publishing, 2014, pp. 678–690.

[39] B. Ellis, *Real-time analytics: Techniques to analyze and visualize streaming data*. John Wiley & Sons, 2014.

[40] M. Barlow, *Real-Time Big Data Analytics: Emergining Architecture*. 2013.

[41] R. Dijkman, S. Peters, and A. ter Hofstede, “A Toolkit for Streaming Process Data Analysis,” in *Enterprise Distributed Object Computing Workshop (EDOCW), 2016 IEEE 20th International*, 2016, pp. 1–9.

[42] M. Stonebraker, U. Çetintemel, and S. Zdonik, “The 8 requirements of real-time stream processing,” *ACM SIGMOD Rec.*, vol. 34, no. 4, pp. 42–47, 2005.

[43] N. Marz and J. Warren, *Big Data: Principles and best practices of scalable realtime data systems*. Manning Publications Co., 2015.

[44] B. Ellis, *Real-time Analytics: Techniques to Analyze and Visualize Streaming Data*. 2014.

[45] S. Kumaresan, S. S. Srinivas, A. Maitra, and N. Kuntagod, “Rapid mHealth-a mobile healthcare application development framework,” *Int. J. Telemed. Clin. Pract.*, vol. 2, no. 1, pp. 42–62,

2017.

[46] S. Mazumder, “Big Data Tools and Platforms,” in *Big Data Concepts, Theories, and Applications*, Springer, 2016, pp. 29–128.

[47] S. Mtapuri-Zinyowera *et al.*, “Evaluation of the PIMA point-of-care CD4 analyzer in VCT clinics in Zimbabwe,” *JAIDS J. Acquir. Immune Defic. Syndr.*, vol. 55, no. 1, pp. 1–7, 2010.

[48] N. P. Pai, C. Vadnais, C. Denkinger, N. Engel, and M. Pai, “Point-of-Care Testing for Infectious Diseases: Diversity, Complexity, and Barriers in Low- And Middle-Income Countries,” *PLoS Med.*, vol. 9, no. 9, p. e1001306, Sep. 2012.