The Road to the Future of Healthcare: Transmitting Interoperable Healthcare Data Through a 5G Based Communication Platform

Argyro Mavrogiorgou[1(✉)], Athanasios Kiourtis[1], Marios Touloupou[1], Evgenia Kapassa[1], Dimosthenis Kyriazis[1], and Marinos Themistocleous[1,2]

[1] Department of Digital Systems, University of Piraeus, Piraeus, Greece
{margy, kiourtis, mtouloup, ekapassa, dimos, mthemist}@unipi.gr
[2] University of Nicosia, Nicosia, Cyprus
themistocleous.m@unic.ac.cy

Abstract. Current devices and sensors have revolutionized our daily lives, with the healthcare domain exploring and adapting new technologies. The rapid explosion of digital healthcare happened with the help of current 4G LTE technologies including innovations such as the continuous monitoring of patient vitals, teleporting doctors to a virtual environment or leveraging Artificial Intelligence to generate new medical insights. The arised problem is that current 4G LTE based communication platforms will not be able to keep up with the exploding connectivity demands. This is where the new 5G technology comes, expected to support ultra-reliable, low-latency and massive data communications. In this paper, an end-to-end approach is being provided in the healthcare domain for gathering medical data, anonymizing it, cleaning it, making it interoperable, and finally storing it through 5G network technologies, for their transmission to a different location, supporting real-time results and decision-making.

Keywords: 5G network · Data integration · Data anonymization
Data cleaning · Data quality · Data interoperability · Healthcare

1 Introduction

In recent years, there has been a lot of focus on how medical and health-monitoring devices, clinical wearables, and remote sensors can contribute to better health for patients and a more efficient healthcare that can drive better systems, population, and patient outcomes [1]. Currently, healthcare is one of the fastest industries to adopt the Internet of Things (IoT) technologies, which help in personalized services, reducing operating costs, and improving patient care and quality of life. However, for most patients and providers, the vague promises of the IoT has not yet led to dramatic changes in how patients experience healthcare [2].

It is undeniable that what is needed is faster connection speeds that will be transforming the healthcare providers - patients relationship, integrating electronic communications into medical care, which can be achieved through the arrival of the

M. Themistocleous and P. Rupino da Cunha (Eds.): EMCIS 2018, LNBIP 341, pp. 383–401, 2019.
https://doi.org/10.1007/978-3-030-11395-7_30

5G networks [3]. From the comfort of their homes, patients will wear remote medical sensors, transmitting their vital signs to healthcare providers that will allow doctors and caregivers to monitor an array of vitals, dynamically manage treatment plans, and conduct a consult or intervention over webcam. To this context, 5G networks will take this recent medical trend to the next level and provide a significant economic boost to the medical community. According to IHS Market, 5G will enable more than $1 trillion dollars in products and services for the global healthcare sector [4], while by 2020, around 50 billion connected devices and 212 billion connected sensors are expected to be supported by the 5G network [5]. For healthcare, this means the birth of entire digital ecosystems that can aid medical research, diagnose conditions, and provide treatment at ever-increasing rates. Hence, 5G represents a completely new way for accomplishing digital networking and upgrading the healthcare experiences, by delivering a holistic personalized view of the patients anytime and anywhere.

However, apart from this challenge, additional problems remain concerning the transmission of the medical data, since 5G networks are providing solutions on 'how' data will be transmitted, and not on 'what' kind of data will be transmitted. Consequently, the problem is not only the difficulty of data exchange between systems, but also the devices' data incompatibility. More particularly, IoT medical devices are typically characterized by a high degree of heterogeneity, in terms of having different capabilities, functionalities, etc. In such a scenario, it is necessary to provide abstractions of these heterogeneous devices and manage their interoperability so as to finally collect medical data out of them [6]. However, existing integration technologies lack of sufficient flexibility to adapt to these changes, as their techniques are both static and sensitive to new or changing device implementations [7].

Even if some researches have overcome this problem, the next problem that arises is that the collected data is difficult to be anonymized due to its inherent heterogeneity, and therefore preventing the sharing of data for secondary purposes (e.g. data analysis, research). At the same time, anonymization and pseudo-anonymization techniques have been heavily debated in the ongoing reform of EU data protection law [8]. Thus, the main question that arises is how to implement anonymization in such a way that will protect individual privacy, but will still ensure that the data is of sufficient quality [9].

Nevertheless, the problem does not stop there. Even if it has become feasible to manage thousands of heterogeneous IoT devices, collect data out of them, and anonymize it, the quality of these devices, as well as their derived data are of dubious quality. Henceforth, the next challenge that emerges is the identification of the devices' quality levels, in conjunction with their derived data that need to be qualitative in the maximum degree. The quality evaluation of the devices, as well as of their produced data are mainly treated as black boxes in the IoT domain, and not much thought is given to their quality when integrated into larger systems [10]. Using such devices without proper quality evaluations may have serious implications in the health domain, whilst the absence of data quality could reduce the grade of the successful interpretation of the out coming results and findings [11].

On top of all these, data heterogeneity is one of the most fundamental challenges in the healthcare domain, as medical devices are rapidly expanding, producing tons of heterogeneous data. In this context, interoperability is the only sustainable way to

enable healthcare entities acting in various locations, and using distinct information systems from different vendors, to collaborate and deliver quality healthcare. A study estimated that savings of approximately $78 billion could be achieved annually if data exchange standards were utilized across the healthcare sector [12]. Multi-site healthcare provisioning and research requires electronic health records (EHRs) data to be restructured into a common format and standard terminologies, linked to other data sources, which is currently delivered through the HL7 FHIR standard [13].

Taking into consideration all of the aforementioned challenges, in this paper an end-to-end approach is being introduced in the healthcare domain for gathering medical data, anonymizing it, processing it, making it interoperable, and finally storing it, through 5G network technologies. In short, an approach is proposed for the dynamic integration of both known and unknown heterogeneous medical devices during runtime, by providing a Dynamic Data Acquisition API for efficiently collecting their data. In order to anonymize this data, an anonymization part is added to the approach, by implementing k-Anonymity techniques for impeding re-identification, and removing some information, letting concurrently the data to be intact for future use, protecting both individual privacy, and making sure that the data is of sufficient quality.

On top of this, in order to assess the quality of the selected heterogeneous devices in conjunction with their derived data, the proposed approach facilitates the devices' reliability, in combination with the quality estimation of their provided data, by firstly cleaning all the acquired data. As soon as the devices' reliability is being completed, and as a result only the reliable devices are kept connected to the platform in conjunction with their corresponding gathered cleaned data, the interoperability of the latter occurs. For that reason, a filtering mechanism is proposed for defining EHRs and medical data as ontologies, which are used to provide a semantic model for representing definition rules of multiple medical standards that are being finally transformed into HL7 FHIR format. All the aforementioned, are being performed through the implementation of a 5G communication network, as well as an enhanced 5G platform with fully virtualized infrastructure, that are likely to change the way that personalized healthcare is currently provided, for both patients and caregivers.

The rest of this paper is organized as follows. Section 2 presents the state of the art regarding the related work in the healthcare context with regards to the 5G networks, data transmission, security, devices and data heterogeneity comparing them with our approach. Section 3 describes the proposed approach of the interoperable data transmission through the proposed eHealth 5G platform, while Sect. 4 analyzes our conclusions and future goals.

2 Related Work

2.1 5G Networks

While many things on the road to 5G are uncertain, it is easy to envision the emergence of new and innovative use cases. This new technology allows a significantly higher data capacity and extremely fast response times, opening up completely new potential applications for a fully connected society. Especially in the healthcare domain this

constitutes a prerequisite, as faster and more accurate results are needed. Consequently, the industry is facing a new wave of digitalization, referred as Healthcare 4.0 [14]. Healthcare 4.0 is a vision of care delivery that is distributed and patient-centered, and there is already evidence of a shift towards virtualization and individualization of care. Virtualization in the healthcare domain comes with the emergence of next generation mobile network strategies (5G) as foundation, in order to complete the transition to personalized care [15]. The delivery of such virtualized care needs to be executed in real time and based on real time data collection, which can be delivered anywhere, anyhow and at any time. Thus, 5G will be a catalyst to trigger innovation of new products and services in the health care domain, by integrating networking, computing and storage resources into one unified infrastructure.

As the 5G Infrastructure Public Private Partnership (5G-PPP) emphasized in [16], the new 5G network should facilitate the integration with the service layer and enable an effective network resource negotiation (i.e. QoS, latency, speed, reliability). Especially in the healthcare domain, various researches have been conducted trying to cover the different aspects of 5G. In more details, in [17] a summary of the benefits offered by 5G to eHealth is presented, pointing out the new imaging techniques and the possibility of a second opinion thanks to the high-speed transmission of X-rays or scans, the telemonitoring that helps to obtain better diagnostics, and the data mining applied to medical data that helps to adjust the treatment among others. Also, an architecture with 5G for a typical Wireless Body Area Network was presented in [18], while in [19] the 5G-Health is introduced as the next generation of eHealth, discussing the possibilities of medical video streaming, thanks to the high speed reached in 5G networks. To that concept, [20] described how 5G technologies will enable new ways of instant exchange of information in order to deliver personalized healthcare data in real time, as well as how to provide more effective and efficient therapeutic approaches. Additional research included in [21], where the authors introduced systems of wearable medical devices and sensors for monitoring physiological recorded signals, within a 5G infrastructure. Finally, in [22] a potential 5G network and machine-to-machine communication is presented for developing and evolving mobile health applications.

2.2 Data Integration

Data integration is considered a key component and, especially in the healthcare domain, where in most of the cases it is considered as a prerequisite in nearly every systematic attempt to achieve integrated care. In the context of healthcare, data integration is a complex process of combining multiple types of data from different heterogeneous sources into a single system/platform [23]. Henceforth, regardless of the way in which devices are connected to each platform, they should be able to be uniformly discoverable and integrated with different platforms, in order for the latter to have access to the sources' medical data.

To this concept, various IoT infrastructures have been proposed in the literature, especially in the healthcare domain, putting their efforts on the integration of heterogeneous medical devices in order to be interoperable and pluggable to different platforms, while offering their data. In more details, the authors in [24] proposed a system to automate the process of collecting patient's vital data via a network of sensors

connected to legacy medical devices and deliver this information to the medical center's cloud for storage, processing, and distribution. Moreover, the authors in [25] proposed an ontology-based cognitive computing eHealth system, aiming to provide semantic interoperability among heterogeneous IoT fitness devices and wellness appliances in order to facilitate data integration, sharing and analysis. In the same notion, in [26] the ContQuest was proposed, a framework that among its functionalities, defined a development process for integrating new data sources including their data description and annotation, by using the Ontology Web Language (OWL) [27] to model and describe data sources. In the same concept, the proposed approaches in [28–31] coped with the frequent modification of data source's schemas, by providing homogeneous views of various data sources based on a domain ontology [7]. In addition, the authors in [32] presented an ontology based on data integration architecture within the context of the ACGT project, where emphasis was given to resolve syntactic and semantic heterogeneities when accessing integrated data sources. Finally, the authors in [33] proposed an IoT based Semantic Interoperability Model (IoT-SIM) to provide semantic interoperability among heterogeneous IoT devices in healthcare domain.

2.3 Data Anonymization

In the healthcare domain, privacy issues must be taken into consideration, as eHealth services offer efficient exchange of the patients' data between different entities [34]. Hence, all this medical data that is exchanged and shared among them must be fully anonymized, overcoming the various security issues that may arise. Therefore, in order to comply with these issues, healthcare stakeholders seek to use personal data protection solutions, using mainly data anonymization [35]. More particularly, data anonymization refers to the process of modifying personal data in such a way that individuals cannot be re-identified and no information about them can be learned [36], ensuring that even if anonymous data is stolen, it cannot be used in violation of the law. Especially in the healthcare domain, all the data that can identify a patient must be removed together with any other information, which in conjunction with other data held by or disclosed to the recipient, could identify the patient [37].

Hence, in order to achieve this kind of anonymization, k-anonymity [38] is most widely implemented, ensuring that each record in a dataset has at least k-1 indistinguishable records. To this context, various researches have been conducted, focusing mainly on data privacy preserving in cloud networks [39, 40], while most of them are mainly using k-anonymity. Apart from this, the authors in [41] adopted the (a, k)-anonymity model as a privacy detection scheme to collect data and propose a new privacy preserving data collection method based on anonymity for healthcare services. Moreover, in order to avoid privacy leakage, the authors in [42] adopted k-anonymity to protect data from re-identification, proposing a semantic-based linkage k-anonymity (LA) to de-identify record linkage with fewer generalizations and eliminate inference disclosure through semantic reasoning, whilst the authors in [43] proposed the LA through which only obfuscated individuals in a released linkage set are required to be indistinguishable from at least k-1 other individuals in the local dataset.

2.4 Data Cleaning

Data cleaning plays a significant role in a broad variety of scientific areas, being responsible for detecting and removing errors and inconsistencies from data, improving its quality [44]. Therefore, data cleaning routines shall be applied to clean the data by filling in missing attributes and values, smoothing and leveling noisy data, identifying and removing outliers, as well as determining and settling inconsistencies [45]. Thus, preparing and cleaning data prior to analysis is a perennial challenge in data analytics, and especially in the healthcare domain, where the produced data is of major importance given that they drive medical decision making.

For that reason, over the last two decades data cleaning has been a key area of research, and many authors have proposed algorithms for data cleaning to remove inconsistencies and noises out of data. The most common inconsistency type has to do with the missing data, for which various algorithms have been proposed so far (i.e. constant substitution, mean attribute value substitution, random attribute value substitution [46]). However, apart from these approaches, there have been proposed several other solutions regarding the different data cleaning problems that may occur. In [47] a method is implemented for managing data duplications, where duplication detection is done either by detecting duplicate records in a single database or by detecting duplicate records in multiple other databases. In the same concept, in [48] a two-step technique that matches different tuples to identify duplicates and merge the duplicate tuples into one is proposed. What is more, to compensate the complexity of data expression, many data cleaning methods are using heuristic rules and user guidance, such as [49–52], which require manual labor for the cleaning process. Hence, in [53] an ontology-based data cleaning solution is implemented, using existing technologies to understand and differentiate the contents of the data, and performing data cleaning without the need of human supervision. Apart from these, in [54] the authors proposed a solution for detecting and repairing dirty data, by offering a commodity data cleaning system that resolves errors like inconsistency, accuracy, and redundancy, by treating multiple types of quality rules holistically. In this context, in [55] a rule-based data cleaning technique is proposed, whereby a set of domain specific rules define how data should be cleaned.

2.5 Sources Reliability

A great attention has been given to the reliability challenge, confronting system reliability as a fundamental requirement of IoT devices. In more details, reliability is a measure of the ability that a system operates as expected under predefined conditions for a predefined time [56]. According to [57] reliability is a technical effort made to ensure that a developed system is free from any fault that can result to failure during operation. It entails that the system is highly dependable and functions maximally at any given time or condition over the period it is created or developed to serve.

To this context, various reliability methods have been proposed in the literature regarding the IoT world, and especially the healthcare domain, putting their efforts on measuring the reliability of the IoT devices that are being used for various health purposes. More particularly, in [58] the authors presented a new methodology for

estimating hardware and software reliability given uncertain use conditions, so as to derive probabilistic estimates for overall system reliability. In the same notion, in [59] a probability-based concept is proposed for measuring the reliability of IoT devices, investigating the proposed model from the perspectives of consumer world, by using things link analysis. Furthermore, in [60] the evaluation of the inter-device reliability of activity monitors was discussed, while in the same concept, in [61] the authors examined the reliability of consumer activity trackers for measuring step count in both laboratory and free-living conditions. The study in [62] evaluated the criterion-related reliability of field-based leg stiffness devices in different testing approaches, by measuring the coefficient of variation, the intraclass correlation coefficient, and the standard error of measurement of these devices. Moreover, in [63] the intra and interrater reliability were evaluated upon the point-of-care nerve conduction device in patients with diabetes and a broad spectrum of nerve injury, while in [64] several criteria and methods presented for assessing reliability of medical equipment.

2.6 Data Interoperability

Interoperability is considered a necessity in electronic healthcare systems. At the same time, the development of medical standards has significantly evolved, yet bearing unsolved challenges with clinical data distributed among heterogeneous sources [65]. The Health Level Seven International (HL7) organization provides the development and the framework of standards, of which the most commonly used is the HL7 v2.x [66], however, HL7 FHIR [13] is the latest standard created by the HL7 organization for the exchange of clinical information, whose main motivation was to simplify and reduce the complexity of the mechanisms and structures defined by it, avoiding the mistakes made in its previous standards (HL7 v3 [67], CDA [68]).

To this context, various researches have been developed for covering the different standards that exist for confronting data interoperability. In more details, the Detailed Clinical Models (DCM) [69] have been used for defining clinical information independently of a specific clinical standard, but aiming to offer the possibility of being transformed into other medical standards. Another approach of data harmonization was the 5-year strategy of NHS Wales focusing on developing an open platform across a fully integrated electronic patient record with the core of the clinical terminology Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) [70]. A process based on HL7 standard and SNOMED CT vocabulary from the biomedical domain and the latest semantic web technologies has been developed and tested within the framework of EURECA EU research project [71], aiming to homogenize the representation and normalization of clinical data. Moreover, in [72] the interoperability among different healthcare systems was reached by annotating the Web Service messages through archetypes defined in OWL, whereas the same researchers presented an approach [73] based on archetypes, ontologies and semantic techniques for the interoperability between HL7 CDA and ISO 13606 systems, which were represented in OWL. Finally, the work of [74] must be mentioned, where the authors presented a solution based on the Enterprise Service Bus that was translated into the healthcare domain using the ideals of HL7 V3 and SNOMED CT.

Taking into consideration all the aforementioned approaches that have been proposed for dealing with the different challenges that exist concerning 5G networks, data integration, anonymization, cleaning, reliability, as well as interoperability in the healthcare domain, we can conclude that our approach is extremely innovative. More particularly, compared with the existing 5G platforms, the proposed eHealth 5G Platform provides simple and powerful access to the medical devices, along with high system capacity, great speed and ultra-high reliability. Apart from this, all the researches that have been made for integrating heterogeneous devices, lack of sufficient flexibility and adaptability to solve challenges arisen from dynamically integrating both known and unknown devices during runtime, a scenario that is fully supported by the data integration part that is developed in our approach. Regarding data anonymization, no innovation is being proposed, as we simply make use of the k-anonymity algorithm. Regarding data cleaning, the existing surveys have presented different approaches for it, lacking an end-to-end iterative data cleaning process, a problem that is totally eliminating in our approach, where an end-to-end iterative data cleaning process is implemented, being capable of cleaning data deriving from both known and unknown devices. As for devices reliability, all the researches that have been proposed so far for characterizing devices' reliability are based only upon the devices' reliability itself, without considering data reliability issues, thus not stating a combined approach, which is crucial for any application in the healthcare domain. Henceforth, the devices reliability part of our approach is considered innovative, as it confronts devices' reliability in combination with their derived data quality. Finally, considering data interoperability, several solutions have been proposed enabling the access to the existing medical data for specific clinical organizations, lacking however to be applied to different medical standards and incoming data, thus not providing a generic approach being able to address heterogeneous healthcare data. To address this gap and confront the interoperability issues, our approach includes a generalized mechanism that employs several matching operations to the HL7 FHIR standard.

3 Proposed Approach

In our approach, an innovative mechanism is proposed for gathering medical data from numerous heterogeneous IoT medical devices, anonymizing this data, cleaning it, making it interoperable, and finally storing it through 5G communication technologies. More specifically, the proposed approach consists of the six (6) main stages: (i) 5G Communication Network, (ii) Data Integration, (iii) Data Anonymization, (iv) Data Cleaning, (v) Devices Reliability, and (vi) Data Interoperability, accompanied with Data Storage, as illustrated in Fig. 1.

5G Network. The architecture of the used 5G Network consists of two (2) major steps related to the 5G communication and integration. Initially, the collection of the data takes place at high reliable edge-nodes, in order to allow the connection of the physical world (i.e. biological system) and the virtual world (i.e. 5G infrastructure). In order to achieve this, the deployment of the edge node in each medical device is being connected through a 5G Radio Access Network (5G RAN) [75], enabling finally the

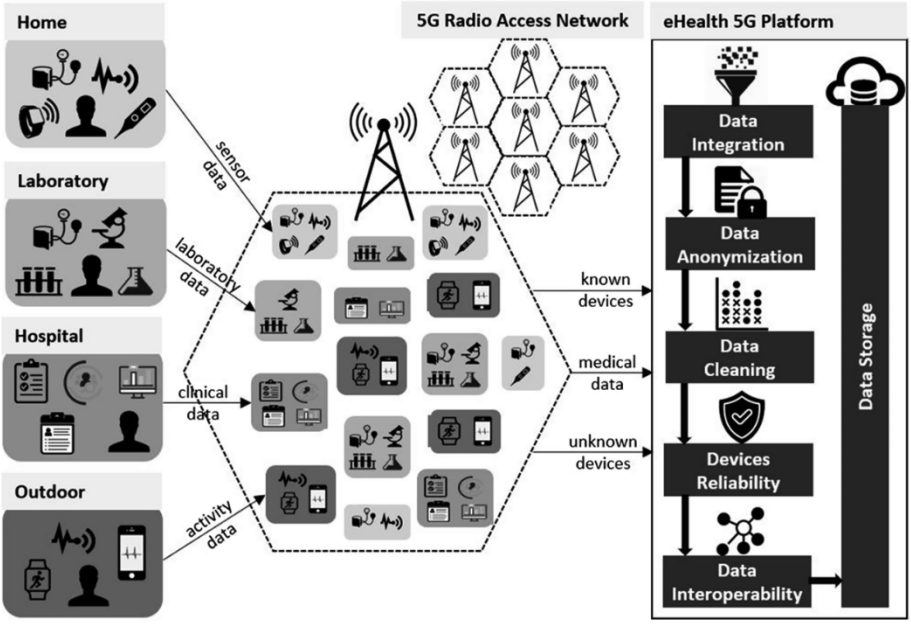

Fig. 1. Architecture of the proposed approach

analysis of the information inside the eHealth 5G Platform through the appropriate integration of different Virtual Network Functions (VNFs) [76].

In more details, in the first step the identification of the available heterogeneous IoT medical devices takes place, through the established 5G RAN communication network. Due to the diverse and extreme requirements of the healthcare data, as well as the eHealth services, the 5G RAN designed to operate in a wide range of spectrum bands, with diverse characteristics, such as channel bandwidth and propagation conditions. The challenge in 5G RANs is how to dynamically assign the foreseen wide range of services with diverse requirements to the many spectrum bands, usage types and radio recourses. Therefore, the proposed approach comes to resolve this challenge by using the Radio Access Network as a Service (RANaaS) [77], by partially centralizing the functionalities of the RAN depending on the actual needs, as well as the network characteristics, being able to handle huge amounts of data, in high-speed with low-cost, providing on-demand resource provisioning delay-aware storage, and high network capacity wherever and whenever needed. Thus, through this established connection, all the data of the connected health devices are being gathered, containing information about the used devices' APIs (i.e. source code) that is assumed that is written in the same programming language, accompanied with the devices' specifications (i.e. hardware and software) that contain the same semantics in terms of specifications' descriptions and measurement units. To this end, it should be noted that the gathered data comes from different entities and systems, referring that is coming either from (i) medical devices that are used by the patients for their in-home monitoring, or (ii) medical devices, EHRs, PHRs that are used by the patients and the healthcare

professionals in medical laboratories for keeping patients' measurements, or (iii) medical devices, EHRs, PHRs that are used by the patients and the healthcare professionals in hospitals for recording and keeping patients' measurements, or finally (iv) medical devices that are used by the patients for their outdoor activities.

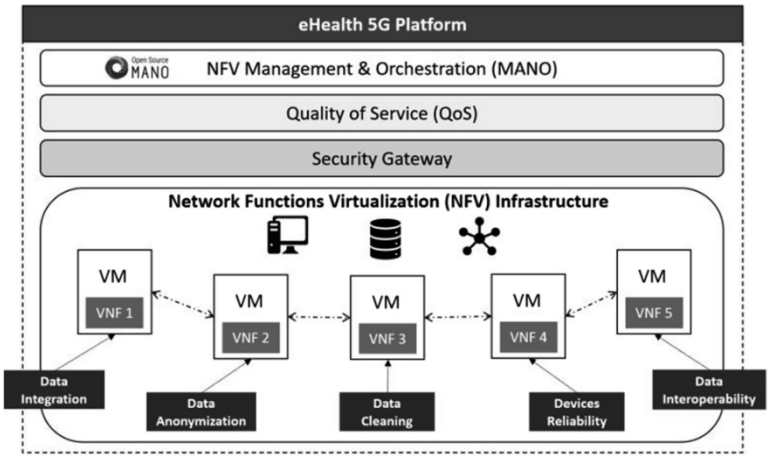

Fig. 2. 5G enhanced platform

As soon as all this data has been collected, the second step occurs, where the 5G architecture of the eHealth platform is being developed, using the technologies of Network Function Virtualization (NFV) [78] and Software Defined Network (SDN) [79]. The value of the 5G SDN (especially in conjunction with NFV and virtualized networking) is its ability to provide network virtualization, automation and creation of new services over virtual resources, affording an extremely manageable and cost-effective architecture, making it ideal for the dynamic, high-bandwidth nature of eHealth. Furthermore, VNFs move individual network functions out of dedicated hardware devices into software that runs on commodity hardware, while it is worth saying that VNFs can run as virtual machines (VMs). In the current approach, we adapted the ongoing 5GTANGO's Service Platform [80], which consists mainly of three (3) discrete blocks: (i) the Service Development Kit, (ii) the Validation and Verification, and (iii) the Service Platform that will be parameterized in our approach. As shown in Fig. 2, the proposed eHealth 5G platform consists of several components which support the whole lifecycle of the VNFs, by the time that they are developed until the time that they are instantiated. Initially, through the NFV Infrastructure all the data management mechanisms that are provided through the eHealth 5G platform (i.e. data integration, data anonymization, etc.) are constructed in the form of VNFs. This transformation is a prerequisite for the efficient operation of the platform, providing it with quite flexibility, cost-efficiency, and scalability, being able to be virtualized in different eHealth platforms running in different entities (i.e. hospitals, health clinics etc.). As soon as all the mechanisms are being transformed into VNFs, the Security

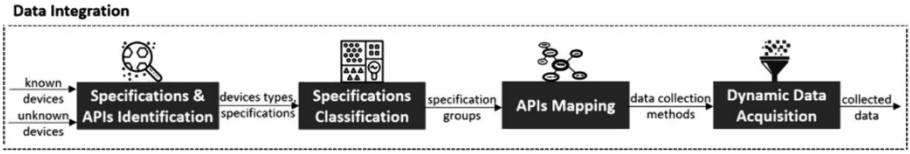

Fig. 3. Data integration process

Gateway component is being implemented, which is responsible for controlling the privileges and the users' access to the platform, by validating the corresponding requests. This is of crucial importance in the context of open 5G ecosystems, where entry barriers are disruptively lowered, without decreasing security [81]. Apart from this, it is of crucial importance to guarantee Quality of Service (QoS) of the developed eHealth 5G platform. For that reason, we create an integrated view of the healthcare applications, the interconnection infrastructure support, and the operational support with common services. Finally, on top of all these, the eHealth 5G platform provides the NFV Management & Orchestration component that uses the MANO framework [82], which is responsible for managing the lifecycle of all the VNFs requests and instances by orchestrating the available infrastructure.

Data Integration. In this stage the data integration occurs, for easily and rapidly integrating heterogeneous IoT medical devices during runtime, concerning both known and unknown devices, so as to be able to collect data out of them. Therefore, this stage contains the first data mechanism of the eHealth 5G platform, while it consists of four (4) discrete substages (Fig. 3), following the work conducted in [83].

In the first substage of the mechanism, as soon as all the devices have been connected to the eHealth 5G platform, these are categorized into either known (in terms of devices of known type, containing predefined APIs methods) or previously unknown devices (in terms of devices of unknown type, containing undefined APIs methods). Afterwards, through the established 5G network connection, information is gathered concerning both devices' specifications (i.e. hardware and software) and APIs (i.e. multiple methods). Thus, information is gathered about (i) both known and unknown devices' specifications, (ii) known devices' APIs in terms of source code and of what exactly each method in the API represents, and (iii) unknown devices' APIs in terms of source code, as the significance of each method in the API is unknown.

Afterwards, in the second substage the classification of the devices' specifications occurs, following the approach proposed in [84]. By knowing (i) the device type of the known devices as well as their specifications, and (ii) the specifications of the unknown devices, their classification occurs, considering the known devices' types and the similar specifications that all these devices may have with the unknown devices. Based on the classification outcomes, the identification of the unknown devices' type takes place, assuming that the devices with the same specifications are of the same type (e.g. all the spirometers will have approximately the same specifications). As a result, all the devices of unknown type are considered as known.

In the third substage, the mapping of the devices' APIs methods occurs. More particularly, in the first substage of Data Integration, knowledge about known devices'

APIs methods was acquired in terms of source code and of what exactly each method in the API represents. However, with regards to the unknown devices' APIs methods, the acquired knowledge referred only to the source code, as the significance of each method in the API was unknown. Henceforth, in this substage it becomes feasible to map the known devices' APIs methods with those of the unknown devices, by comparing the API methods of the devices of the same type (e.g. all the spirometers). In order to achieve this mapping, for each one of these devices a Generic API Ontology (GAO) is constructed, based on the approaches proposed in [7, 85], in order to identify and model a hierarchical tree of the different classes and sub-classes of the semantics of the devices' APIs methods. In more details, each GAO contains different ontologies for each different method of each device's API, thus a hierarchical tree is being created for each API, allowing us to understand and probabilistically map the similar methods. In our case, the mechanism has to identify and map the method that is responsible for gathering the unknown devices' data, thus the method that has been assigned with higher probability levels, is automatically assigned as the most appropriate method.

Finally, as soon as this mapping is completed, in the fourth substage the implementation of the Dynamic Data Acquisition API occurs. More particularly, the latter constitutes of a unified API that merges into a single unified data method all the different devices APIs' data methods that are responsible for collecting data, and thus the collection of devices' data takes place.

Data Anonymization. In this stage the data anonymization occurs, where the collected data is pre-processed through k-Anonymity using data suppression and data generalization, following two (2) different substages, as depicted in Fig. 4.

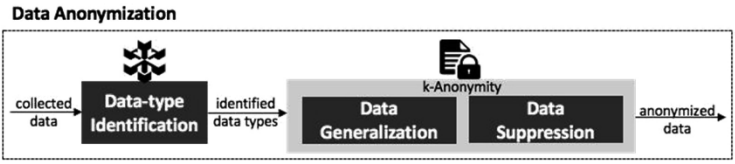

Fig. 4. Data anonymization process

In the first substage, the data-type identification of the collected data takes place, through which we are able to identify whether an individual value of an attribute can be anonymized through the data suppression or the data generalization method, by identifying the data type of each value. It should be mentioned that only the personal data (i.e. data that identify a person) are being filtered through this mechanism.

Therefore, in the second substage, the anonymization of the collected data occurs, where k-Anonymity is being implemented, applying data generalization and/or data suppression, depending on the results of the data-type identification. In more details, through the data suppression method, certain values of the attributes are replaced by a hashtag '#', according to their semantics and to what they represent. Regarding the data generalization method, individual values of attributes are replaced with a broader category, being given a range where the anonymized value can be found in between.

Consequently, implementing the corresponding method upon the collected data, we result into the fully anonymization of it, taking into consideration that the numeric values are being anonymized through the data generalization method, while all the other types of values are anonymized through the data suppression method.

Data Cleaning. In this stage, the cleaning of the anonymized data takes place, which is received as an input in conjunction with the device type that gathered this data, maintaining the data model of each device type. Within this data model, the elements of the data are defined in addition to a set of constraints, predefined rules for the corrective actions, and the automated data filling. Therefore, for each dataset four (4) discrete substages are followed sequentially, as illustrated in Fig. 5.

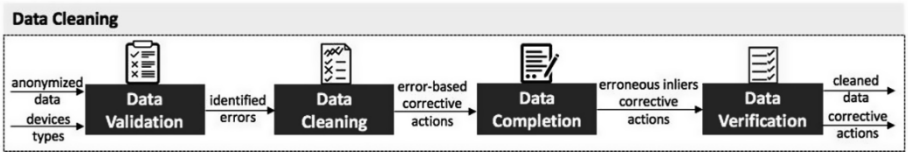

Fig. 5. Data cleaning process

In the first substage, the data validation occurs that identifies all the errors associated with the conformance to a set of predefined rules, such as data type (i.e. integer, string, etc.), range constraints (i.e. minimum and maximum values), uniformity (i.e. data format), predefined values (i.e. values selected from a predefined list), and mandatory fields. Hence, a variety of validity checks is performed aiming to safeguard the accuracy and the consistency of the data by ensuring the conformance both to the specified constraints on the data model for this device type and the identified duplicates.

In the second substage, the data cleaning occurs that eliminates the errors identified in the previous substage, where based on the set of the predefined rules, corrective/removal actions are applied on the identified erroneous records of the data.

Sequentially, in the third substage the data completion takes place that safeguards the appropriateness and completeness of the data, especially referring to erroneous inliers, where the conformance to mandatory fields and required non-empty attributes of the data is ensured based on the predefined conformance rules of the data model.

Finally, in the fourth substage, the data verification occurs that executes the evaluation of the undertaken actions in the previous substages, ensuring the accuracy and consistency of the cleaned data. Thus, the final results are produced, indicating the undertaken total corrective actions in combination with the derived cleaned data.

Devices Reliability. In this stage the devices reliability takes places, in combination with the quality estimation of their provided data, as depicted in Fig. 6. This stage is of major importance, as it is not sufficient to keep all the derived data and use it for further analysis, as many of it may have derived either from unreliable devices, or from reliable devices being uncleaned and faulty. For that reason, it is necessary to measure and evaluate the quality of both the connected devices and their produced data, so as to

finally keep only the reliable data that comes from only reliable devices. In our case, for measuring devices' reliability we captured the metric of the availability of the connected devices, an important metric for assessing the quality of the devices [57]. Therefore, we measure each device's availability by getting the corresponding values, setting a timestamp in order to measure how often each device communicates with the platform and provides its data. However, it is not sufficient to measure only the devices' availability for deciding whether the latter is being considered as reliable or not, but it is more effective to measure also the data quality of these devices. For that reason, we use as an input from the Data Cleaning stage the number of the undertaken actions that were applied upon the collected datasets, in order to correlate it with the availability results of the corresponding devices that produced these datasets, and finally decide whether each device, and as a result its derived data, are considered as reliable or not.

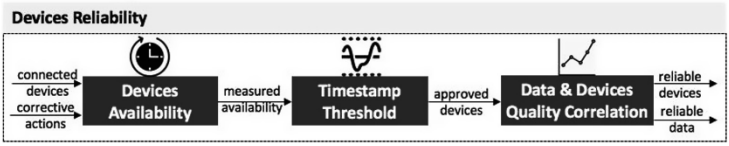

Fig. 6. Devices reliability process

Data Interoperability. In this stage, the final transformation of the data takes place. Thus, the data interoperability process occurs, including an automated way for transforming the ingested data into HL7 FHIR format in terms of structure (Fig. 7).

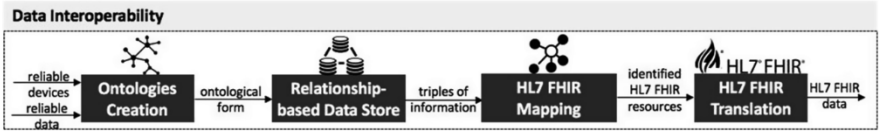

Fig. 7. Data interoperability process

In the first substage, ontologies are created for the source data by transforming the provided data into an ontological form. Thus, this substage delivers the means so that the different relationships, classes and instances are discovered, providing a way for easily classifying these categories, enabling easier manipulation for the next substages.

Afterwards, the second substage provides a relationship-based data store for storing the identified relationships, classes, and instances, making it easier to perform queries through the collected data that can possibly contain information concerning one or more of the stored information. Through this substage it is easier to probabilistically identify faulty or missed relationships among different classes/instances, through a relationship matching mechanism that contains a functionality for identifying missing values, and re-assigning the relationships that have a larger degree of association to a specific class.

Sequentially, the third substage provides a mechanism that offers the capability of understanding and interpreting the semantic meaning of the different classes that have already been stored into the previous substage. Afterwards, this substage is incorporating a mechanism that iterates and scans through the different HL7 FHIR resources, in order to probabilistically map the semantics of the stored classes with a specific HL7 FHIR resource. In the end, the HL7 FHIR resource with higher probability levels of correspondence is automatically assigned to the identified class.

Finally, the fourth substage provides a mechanism for setting the final HL7 FHIR-based form of the classes. Hence, the classes along with their identified HL7 FHIR resources are obtained, including the name of the HL7 FHIR resource along with the specific attribute that the class may belong to, translating them using the HL7 FHIR-based formatting (i.e. *Resource.attribute*). Thus, all the data is translated into a common format, being finally stored into the eHealth 5G platform's database.

4 Conclusions

While current devices have revolutionized our daily lives in multiple domains, the quantity of available healthcare data is rising rapidly, far exceeding the capacity to deliver personal or public health benefits from analyzing this data. Hence, a substantial overhaul of methodology is required to address the real complexity of health. In this paper, an innovative end-to-end approach was proposed for gathering medical data, anonymizing it, cleaning it, making it interoperable, and finally storing it through 5G network technologies. Therefore, it combined core technologies that are crucial in the healthcare domain, for delivering results of high-reliability and efficiency. Even though there have been proposed several techniques for addressing the aforementioned data domains, most of these have been designed to give a solution to specific problems, with low flexibility and adaptability. Contrariwise, our approach promises faster results of high accuracy, merging multiple innovative data manipulation techniques.

Nevertheless, the proposed approach still has to be compared with multiple mechanisms that provide similar services, and evaluated with datasets of different nature and size, and in multiple systems, so as to have better interpretable results. Our future work includes that the mechanism will be also evaluated by testing it with a huge amount of heterogeneous IoT medical devices of different types. We also plan to extend the list of the supported data cleaning constraints including more advanced and sophisticated constraints, while we aim to configure the data anonymization mechanism by testing it with additional data anonymization algorithms. Finally, we plan to evaluate our approach with multiple healthcare data, including formats of unknown nature.

Acknowledgements. A. Mavrogiorgou and A. Kiourtis would like to acknowledge the financial support from the "Hellenic Foundation for Research & Innovations (HFRI)". Moreover, part of this work has been partially supported by the 5GTANGO project, funded by the European Commission under Grant number H2020ICT-2016-2 761493 through the Horizon 2020 and 5G-PPP programs (http://5gtango.eu).

References

1. Population health outcomes. http://www.healthcatalyst.com/population-health-outcomes-3-keys-to-drive-improvement
2. The role of IoT in the healthcare industry. https://hackernoon.com/the-role-of-internet-of-things-in-the-healthcare-industry-759b2a1abe5
3. Healthcare needs 5G. https://www.chilmarkresearch.com/healthcare-needs-5g/
4. How will 5G impact different industries? http://prescouter.com/2018/01/5g-impact-different-industries
5. The Journey to 5G. http://www.healthcareitnews.com/news/journey-5g
6. Pires, F., et al.: A platform for integrating physical devices in the Internet of Things. In: Embedded and Ubiquitous Computing (EUC), pp. 234–241. IEEE (2014)
7. Gong, P.: Dynamic integration of biological data sources using the data concierge. Health Inf. Sci. Syst. **1**, 1–19 (2013)
8. GDPR requirements. https://www.delphix.com/white-paper/gdpr
9. El Emam, K., Arbuckle, L.: Anonymizing Health Data: Case Studies and Methods to get you started, 2nd edn, p. 1005. O'Reilly Media Inc., Newton (2013)
10. Kruger, P., Hancke, G.: Benchmarking internet of things data sources. In: 12th IEEE International Conference on Industrial Informatics (INDIN). IEEE (2014)
11. Macfarlane, S., Tannath, T., Scott, J., Kelly, V.: The validity and reliability of global positioning systems in team sport: a brief review. JSCR **30**(5), 1470–1490 (2016)
12. Mead, C.: Data interchange standards in healthcare IT-computable semantic interoperability. JHIM **20**, 71–78 (2006)
13. HL7 FHIR. https://www.hl7.org/fhir/
14. HEALTHCARE 4.0: A NEW WAY OF LIFE? http://www.vph-institute.org/news/healthcare-4-0-a-new-way-of-life.html
15. A new Generation of eHealth Systems Powered by 5G. http://www.wwrf.ch/files/wwrf/content/files/publications/outlook/Outlook17.pdf
16. 5G on eHealth. https://5g-ppp.eu/wp-content/uploads/2016/02/5G-PPP-White-Paper-on-eHealth-Vertical-Sector.pdf
17. INTERNET OF THINGS & 5G REVOLUTION. http://www.astrid-online.it/static/upload/stud/studio-i-com_internet_5g_.pdf
18. Mishra, A., Agrawal, P.: Continuous health condition monitoring by 24 × 7 sensing and transmission of physiological data over 5G cellular channels. In: ICNC, pp. 584–590 (2015)
19. Banaee, H., et al.: Data mining for wearable sensors in health monitoring systems: a review of recent trends and challenges. Sensors **13**(12), 17472–17500 (2013)
20. Ryan, M., et al.: Facilitating health behaviour change and its maintenance: interventions based on self-determination theory. Eur. Health Psychol. **10**, 2–5 (2008)
21. Oleshchuk, V., Fensli, R.: Remote patient monitoring within a future 5G infrastructure. Wirel. Pers. Commun. **57**, 431–439 (2011)
22. Mattos, W., Gondim, P.: M-health solutions using 5G networks and M2M communications. IT Prof. **18**(3), 24–29 (2016)
23. Leventer-Roberts, M., Balicer, R.: Data integration in health care. In: Amelung, V., Stein, V., Goodwin, N., Balicer, R., Nolte, E., Suter, E. (eds.) Handbook Integrated Care, pp. 121–129. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-56103-5_8
24. Rolim, C.O., et al.: A cloud computing solution for patient's data collection in health care institutions. In: Second International Conference on ETELEMED 2010. IEEE (2010)
25. Carbonaro, A., Piccinini, F., Reda, R.: Integrating heterogeneous data of healthcare devices to enable domain data management. JeLKS **14**(1), 45–56 (2018)

26. Pötter, B., Sztajnberg, A.: Adapting heterogeneous devices into an IoT context-aware infrastructure. In: Software Engineering for Adaptive and Self-Managing, pp. 64–74. ACM (2016)
27. OWL. https://www.w3.org/TR/owl-guide/
28. Globle, C., et al.: Transparent access to multiple bioinformatics information sources. IBM Syst. J. **40**, 534–551 (2001)
29. Donelson, L., et al.: The BioMediator system as a data integration tool to answer diverse biologic queries. In: Proceedings of MedInfo, pp. 768–772 (2004)
30. Philippi, S.: Light-weight integration of molecular biological databases. Bioinformatics **20**, 51–57 (2004)
31. Eckman, B., Lacroix, Z., Raschid, L.: Optimized seamless integration of biomolecular data. In: IEEE International Conference on Bioinformatics and Biomedical Engineering, pp. 23–32 (2001)
32. Martín, L., et al.: Ontology based integration of distributed and heterogeneous data sources in ACGT. In: HEALTHINF, pp. 301–306 (2008)
33. Jabbar, S., et al.: Semantic interoperability in heterogeneous IoT infrastructure for healthcare. Wirel. Commun. Mobile Comput. (2017)
34. Truta, T., Vina, B.: Privacy protection: p-sensitive k-anonymity property. In: 22nd International Conference on Data Engineering Workshops, Atlanta (2006)
35. El Emam, K.: Data anonymization practices in clinical research. a descriptive study. University of Ottawa (2006)
36. El Emam, K., et al.: A systematic review of re-identification attacks on health data. PLoS One **6**(12), e28071 (2011)
37. Zhong, S., et al.: Privacy-enhancing k-anonymization of customer data. In: PODS 2005, pp. 139–147 (2004)
38. Sweeney, L.: k-anonymity: a model for protecting privacy. Int. J. Unc. Fuzz. Knowl. Based Syst. **10**(5), 557–570 (2002)
39. Benjamin, E., et al.: Systematic literature review on the anonymization of high dimensional streaming datasets for health data sharing. Procedia Comput. Sci. **63**, 348–355 (2015)
40. Dubovitskaya, A., Urovi, V., Vasirani, M., Aberer, K., Schumacher, M.I.: A cloud-based eHealth architecture for privacy preserving data integration. In: Federrath, H., Gollmann, D. (eds.) SEC 2015. IAICT, vol. 455, pp. 585–598. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-18467-8_39
41. Li, H., et al.: (a, k)-anonymous scheme for privacy-preserving data collection in IoT-based healthcare services systems. J. Med. Syst. **42**(3), 56 (2018)
42. Lu, Y., Sinnott, R.O., Verspoor, K.: A semantic-based k-anonymity scheme for health record linkage. Stud. Health Technol. Inform. **239**, 84–90 (2017)
43. Lu, Y., Verspoor, K., Sinnott, R.O., Parampalli, U.: Effective preservation of privacy during record linkage. In: School of Computing and Information Systems, p. 25 (2017)
44. Fatima, A., Nazir, N., Gufran, K.: Data cleaning in data warehouse: a survey of data pre-processing techniques and tools. JITCS **9**, 50–61 (2017)
45. Rahm, E., Do, H.: Data cleaning: problems and current approaches. IEEE Bull. Tech. Comm. Data Eng. **23**(4), 2000–2012 (2000)
46. Krishnan, S., Haas, D., Franklin, M., Wu, E.: Towards reliable interactive data cleaning: a user survey and recommendations. In: HILDA, California (2016)
47. Dallachiesa, M., et al.: NADEEF: a commodity data cleaning system. In: ACM SIGMOD International Conference on Management of Data, New York (2013)
48. Dagade, A., Mali, M., Pathak, N.: Survey of data duplication detection and elimination in domain dependent and domain-independent databases. IJARCSMS **4**(5), 238–243 (2016)

49. Benjelloun, O., et al.: Swoosh: A Generic Approach to Entity Resolution. Stanford InfoLab, Stanford (2005)
50. Bohannon, P., Fan, W., Flaster, M., Rastogi, R.: A cost-based model and effective heuristic for repairing constraints by value modification. In: ACM SIGMOD (2005)
51. Cong, G., Fan, W., Geerts, G., Jia, X., Ma, S.: Improving data quality: consistency and accuracy. In: The 33rd International Conference on Very Large Data Bases, Vienna (2007)
52. Fan, W., et al.: Towards certain fixes with editing rules and master data. VLDB J. **21**(2), 213–238 (2012)
53. Yakout, M., et al.: Guided data repair. Proc. VLDB Endowment **4**(5), 279–289 (2011)
54. Cheng, K., Hong, J.: A novel data cleaning with data matching. Adv. Sci. Technol. Lett. **136**, 161–169 (2016)
55. Gohel, A., et al.: A commodity data cleaning system. Int. Res. J. Eng. Technol. **4**(5), 1011–1014 (2017)
56. Joseph, W.: Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. J. Strength Cond. Res. **19**(1), 231 (2005)
57. Toporkov, A.: Criteria and methods for assessing reliability of medical equipment. Biomed. Eng. **42**(1), 11–16 (2008)
58. Mudasir, A.: Reliability models for the internet of things: a paradigm shift. In: IEEE International Symposium on ISSREW. IEEE (2014)
59. Zin, T.T., et al.: Reliability and availability measures for Internet of Things consumer world perspectives. In: 5th Global Conference on Consumer Electronics. IEEE (2016)
60. Ryan, R., et al.: Validity and reliability of Fitbit activity monitors compared to ActiGraph GT3X+ with female adults in a free-living environment. J. Sci. Med. Sport **20**(6), 578–582 (2017)
61. Kooiman, T., et al.: Reliability and validity of ten consumer activity trackers. BMC Sport. Sci. Med. Rehabil. **7**(1), 24 (2015)
62. Ruggiero, L., et al.: Validity and reliability of two field-based leg stiffness devices: implications for practical use. J. Appl. Biomech. **32**(4), 415–419 (2016)
63. Justin, L., et al.: Reliability and validity of a point-of-care sural nerve conduction device for identification of diabetic neuropathy. PLoS One **9**(1), e86515 (2014)
64. Misra, P., et al.: An interoperable realization of smart cities with plug and play based device management (2015)
65. Rastegar-Mojarad, M., et al.: Need of informatics in designing interoperable clinical registries. Int. J. Med. Inform. **108**, 78–84 (2017)
66. Introduction to HL7 Standards. http://www.hl7.org/implement/standards/
67. HL7 v3. https://www.hl7.org/fhir/comparison-v3.html
68. The HL7 Clinical Document Architecture. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC130066/
69. Goossen, W., et al.: Detailed clinical models. Healthc. Inform. **16**, 201–214 (2010)
70. Wardle, M., Spencer, A.: Implementation of SNOMED CT in an online clinical database. Futur. Hosp. J. **4**(2), 126–130 (2017)
71. EURECA EU project. https://www.dceureca.eu/
72. Dogac, A., et al.: Artemis: deploying semantically enriched web services in the healthcare domain. Inf. Syst. **31**, 321–339 (2006)
73. Schulz, S., Udo, H.: Part-whole representation and reasoning in formal biomedical ontologies. AI Med. **34**(3), 179–200 (2005)
74. Ryan, A., Eklund, P.: A framework for semantic interoperability in healthcare. Stud. Health Tech Inform. **136**, 759 (2008)
75. Marsch, P., et al.: 5G radio access network architecture: design guidelines and key considerations. IEEE Commun. Mag. **54**(11), 24–32 (2016)

76. VNF. https://searchsdn.techtarget.com/definition/virtual-network-functions
77. Ferreira, L., et al.: An architecture to offer cloud-based radio access network as a service. In: European Conference on Networks and Communications. IEEE (2014)
78. Network Functions Virtualisation. http://www.etsi.org/technologies-clusters/technologies/nfv
79. SDN. https://www.opennetworking.org/sdn-definition/
80. 5G Development and Validation Platform for global Industry-specific Network Services and Apps. http://5gtango.eu/
81. Parada, C., et al.: 5GTANGO: A Beyond-MANO Service Platform (in press)
82. Open Source MANO. http://www.etsi.org/technologies-clusters/technologies/nfv/open-source-mano
83. Mavrogiorgou, A., Kiourtis, A., Kyriazis, D.: Plug'n'play IoT devices: an approach for dynamic data acquisition from unknown heterogeneous devices. In: Barolli, L., Terzo, O. (eds.) CISIS 2017. AISC, vol. 611, pp. 885–895. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-61566-0_84
84. Mavrogiorgou, A., Kiourtis, A., Kyriazis, D.: A comparative study of classification techniques for managing IoT devices of common specifications. In: Pham, C., Altmann, J., Bañares, J.Á. (eds.) GECON 2017. LNCS, vol. 10537, pp. 67–77. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-68066-8_6
85. Kiourtis, A., et al.: Aggregating heterogeneous health data through an ontological common health language. In: DeSE 10th International Conference. IEEE (2017)