System model for integration of wearable smart device data into a central health information system

Koren, Ana

University of Zagreb
FACULTY OF ELECTRICAL ENGINEERING AND COMPUTING

Ana Koren

SYSTEM MODEL FOR INTEGRATION OF WEARABLE SMART DEVICE DATA INTO A CENTRAL HEALTH INFORMATION SYSTEM

DOCTORAL THESIS

Zagreb, 2023.

University of Zagreb
FACULTY OF ELECTRICAL ENGINEERING AND COMPUTING

Ana Koren

SYSTEM MODEL FOR INTEGRATION OF WEARABLE SMART DEVICE DATA INTO A CENTRAL HEALTH INFORMATION SYSTEM

DOCTORAL THESIS

Supervisor:
Associate Professor Marko Jurčević

Zagreb, 2023.

Sveučilište u Zagrebu
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

Ana Koren

MODEL SUSTAVA ZA PRIHVAT MJERNIH PODATAKA NOSIVIH PAMETNIH UREĐAJA U SREDIŠNJI ZDRAVSTVENI INFORMACIJSKI SUSTAV

DOKTORSKI RAD

Mentor:
izv. prof. dr. sc. Marko Jurčević

Zagreb, 2023.

Doktorski rad izrađen je na Sveučilištu u Zagrebu Fakultetu elektrotehnike i računarstva, na Zavodu za osnove elektrotehnike i električka mjerenja

Mentor: izv. prof. dr. sc. Marko Jurčević

Doktorski rad ima: 149 stranica

Doktorski rad br.:_____

About supervisor

Marko Jurčević was born in Zagreb, Croatia, in 1980. He received the B.Sc. degree in information and communication technology and Ph.D. in electrical engineering with the Department of Electrical Engineering Basics and Measurements, Faculty of Electrical Engineering and Computing, University of Zagreb, in 2003 and 2010 respectively. He is currently an Associate Professor at the University of Zagreb. His research interests include virtual instrumentation and design and application of the remote measurements and calibration for electrical quantities and information security of measurement and automatization systems.

O mentoru

Marko Jurčević rođen je 1980. godine u Zagrebu, Hrvatska. Diplomirao je informacijske i komunikacijske tehnologije, a doktorirao elektrotehniku na Zavodu za osnove elektrotehnike i mjerenja Fakulteta elektrotehnike i računarstva Sveučilišta u Zagrebu 2003. i 2010. godine. Trenutno je izvanredni profesor na Sveučilištu u Zagrebu. Njegovi istraživački interesi uključuju virtualnu instrumentaciju te dizajn i primjenu daljinskih mjerenja i kalibracije za električne veličine i informacijsku sigurnost mjernih i automatizacijskih sustava.

Summary

The area of research is related to the issue of including personalized health data collected via the Internet of Medical Things (IoMT) devices, such as personal tracking devices, into the Electronic Health Record (EHR) that today constitutes part of the central health information systems of most European countries.

In order for the collected data to be used in EHR, it is necessary to guarantee its quality and inviolability. Therefore, data cleaning and processing is necessary. As part of the scientific work, a comparison of different models of health data cleaning has been performed. A method for optimizing selected models of data cleaning and data transformation with the aim of improving accuracy and precision has been proposed.

Furthermore, data compliance with existing standards and regulations must be ensured, which has been achieved by defining semantic constraints and validation processes via a schematron, a structural schema expressed in XML.

Finally, these components were integrated as modules into a central health system model that would allow the use of IoMT data in EHR.

Keywords: Internet of Medical Things, Electronic Health Records, sensors, standardization

Model sustava za prihvat mjernih podataka nosivih pametnih uređaja u središnji zdravstveni informacijski sustav

Senzori su izvediv i praktičan način za nenametljivo i kontinuirano praćenje vitalnih znakova. Dakle, korištenje senzora u medicinskom području potencijalno može pružiti značajnu pomoć u pronalaženju uzroka bolesti ili postavljanju dijagnoze u složenim slučajevima. Potencijal korištenja zdravstvenih podataka prikupljenih od strane korisnika je ogroman, što ga čini paradigmom koja obećava poboljšanje zdravlja i dobrobiti ljudi. Nosivi senzori su sveprisutni jer se mogu integrirati u ručni sat, narukvicu, ljepljivu traku ili odjeću. Ovi senzori prikupljaju osobne zdravstvene podatke, kao što su broj otkucaja srca, krvni tlak, tjelesna aktivnost i kretanje tijela, temperatura, brzina disanja i zasićenost kisikom. Opseg i raznolikost zdravstvenih podataka doživljavaju eksponencijalni rast s usvajanjem elektroničkih zdravstvenih zapisa, medicinskih nosivih uređaja i osobnih uređaja za praćenje; posljedično, korištenje ovih podataka do njihovog punog potencijala dalo bi nam bolje razumijevanje i olakšalo optimalne kliničke odluke. Ove bi informacije bile od iznimne pomoći zdravstvenim djelatnicima jer bi im kontinuirano pružale vitalan, dubinski uvid u stanje pacijenata, omogućujući poboljšani, individualiziraniji pristup njezi pacijenata. Nadalje, takva promjena perspektive na više proaktivne, umjesto reaktivne medicinske usluge može dovesti do ukupnog pada troškova javnog zdravstva. Cilj studije bio je istražiti bežične senzore u kontekstu Interneta stvari s ciljem predstavljanja modela rješenja za njihovu upotrebu u javnom zdravstvenom sustavu.

Poglavlje 2 zadire u svijet elektroničkih zdravstvenih zapisa (EHR), počevši s opsežnom definicijom. EHR sadrži pacijentovu medicinsku povijest, uključujući prethodne medicinske probleme, susrete i tretmane, alergije, imunizacije, vitalne znakove, laboratorijska izvješća te radiološke i dijagnostičke slike, sve u digitalnom obliku. Budući da mu je svrha pružanje njege pacijentu u više zdravstvenih organizacija, nije u vlasništvu niti jednog subjekta. Stoga mora biti sposoban integrirati podatke iz međusobno neovisnih zdravstvenih sustava i omogućiti njihovu interoperabilnost. Ovo poglavlje istražuje koncept osobnih zdravstvenih kartona (PHR) i prati povijesni razvoj EHR-a. U poglavlju se raspravlja o različitim primjenama i budućem potencijalu EHR-a, naglašavajući njegov značajan utjecaj tijekom pandemije COVID-19. Dalje zadire u ključne standarde i propise koji reguliraju EHR, s fokusom na HL7 FHIR i openEHR.

Poglavlje zaključuje ispitivanjem specifičnog konteksta elektroničkih zdravstvenih kartona u Hrvatskoj i pregledom opće arhitekture sustava EHR-a.

Treće poglavlje istražuje područje senzora i prikupljanje podataka u kontekstu zdravstvene skrbi. Započinje proučavanjem karakteristika senzora, uključujući detaljnu raspravu o klasifikaciji pogrešaka i važnosti kalibracije senzora. Poglavlje zatim prebacuje fokus na ulogu bežičnih senzorskih mreža u zdravstvu, naglašavajući njihov značaj u modernim zdravstvenim sustavima. Značajan dio poglavlja posvećen je nosivim uređajima za praćenje aktivnosti, ističući njihovu rastuću prevalenciju i utjecaj na zdravstvenu skrb. Rasprava se proširuje na kvalitetu podataka (QOD) koju generiraju ovi uređaji za praćenje aktivnosti, bacajući svjetlo na pouzdanost i točnost prikupljenih informacija. Sveukupno, Poglavlje 3 pruža sveobuhvatno ispitivanje tehnologije senzora, bežičnih mreža i nosivih uređaja za praćenje, naglašavajući njihovu ključnu ulogu u krajoliku prikupljanja podataka u domeni zdravstvene skrbi.

U idućem poglavlju, fokus se pomiče na razvoj modela sustava za čišćenje i transformaciju podataka u kontekstu zdravstvene skrbi. Poglavlje počinje istraživanjem modela vođenih podacima posebno dizajniranih za čišćenje podataka senzora eHealtha. Ulazi u zamršenost ovih modela, naglašavajući njihovu primjenu u poboljšanju kvalitete i pouzdanosti zdravstvenih podataka. Značajan dio poglavlja posvećen je praktičnom slučaju korištenja EKG signala, pri čemu se uspoređuju različiti modeli. Ova komparativna analiza služi za isticanje prednosti i slabosti različitih pristupa u radu s određenim vrstama zdravstvenih podataka. Poglavlje zatim proširuje svoje istraživanje na poboljšanja modela, posebno u području algoritama klasifikacije, prikazujući načine za poboljšanje učinkovitosti i točnosti procesa čišćenja i transformacije podataka. U sklopu znanstvenog rada napravljena je usporedba različitih modela čišćenja zdravstvenih podataka. Predložena je metoda optimizacije odabranih modela čišćenja podataka te transformacije podataka s ciljem poboljšanja točnosti i preciznosti. Čišćenje podataka ima za cilj kako otkriti i ukloniti pogreške u podacima koje potječu iz početnih podataka. Tehnike čišćenja podataka glavni su fokus nastojanja da se riješi problem čišćenja podataka u velikim WSN-ovima. Pronalaženje visokoučinkovitog modela predviđanja je neophodno jer korištenje modela predviđanja u medicinskoj skrbi zahtijeva veliku preciznost. Brojni modeli uzeti su u obzir, uključujući strojeve potpornih vektora, stabla odlučivanja, slučajne šume i višestruku linearnu regresiju te neuronske mreže. Skup podataka koji se koristi je skup podataka otvorenog pristupa MHEALTH (Mobile HEALTH). Skup podataka MHEALTH sadrži snimke kretanja tijela i vitalnih znakova za deset volontera različitih profila (spol, dob i različita razina fizičke

spremnosti) tijekom bavljenja različitim fizičkim aktivnostima. Sudionikova prsa, desni zglob i lijevi gležanj bili su opremljeni senzorima koji bilježe ubrzanje, brzinu okretanja i orijentaciju magnetskog polja različitih dijelova tijela. Senzor, koji se nalazi na prsima, generira očitanja EKG-a u 2 odvoda, koja se mogu koristiti za rutinsko praćenje rada srca, probir različitih aritmija ili ispitivanje utjecaja tjelesne aktivnosti na EKG. Za sve senzorske modalitete koristi se frekvencija uzorkovanja od 50 Hz, što se smatra dovoljnim za bilježenje ljudske aktivnosti. Utvrđeno je da se ovaj skup podataka generalizira na uobičajene dnevne aktivnosti. (npr. trčanje naspram stajanja). U podacima jednog od rezultata, 10% podataka je odbačeno, a vrijednosti koje nedostaju izračunate su pomoću statističkih tehnika i tehnika strojnog učenja kako bi se usporedila kvaliteta izračuna različitih modela vođenih podacima. Procijenjene vrijednosti se zatim uspoređuju sa stvarnim vrijednostima, a modeli se procjenjuju suprotstavljanjem izračunatih pogrešaka. Pogreške se razlikuju među subjektima, a višestruka linearna regresija i neuronske mreže dosljedno su nadmašivale sve ostale metode. Kao rezultat toga, odlučeno je koristiti te tehnike za daljnje usavršavanje modela. Metoda koristi stabla odlučivanja i slučajne šume za otkrivanje segmenata skupa podataka gdje zapisi unutar segmenta imaju veću sličnost i korelacije atributa kako bi se poboljšali rezultati imputacije. Podaci koji nedostaju zatim se imputiraju korištenjem korelacije i sličnosti. Ova se metoda može koristiti za širok raspon podataka, a kao rezultat toga, ima brojne potencijalne upotrebe. Pretpostavljajući da postoji veća korelacija u dijelovima koji su odvojeni tjelesnim aktivnostima kojima se bavite (kao što je nepomično sjedenje, hodanje ili trčanje), što odražava utvrđeni odnos između tjelesne aktivnosti i srčane aktivnosti. Ovi se podaci ručno unose u MHEALTH skup podataka monitora. Međutim, ako je vrlo precizna kategorizacija segmenta podataka izvediva, ona bi se mogla koristiti za bilo koji usporedivi skup podataka o m-zdravlju. Osim toga, budući da mnogi zdravstveni podaci potječu od pametnih narukvica ili pametnih satova, izračuni koji slijede koncentriraju se na podatke sa senzora koji se nalaze na desnom zapešću, i žiroskop i ubrzanje. Posljednji korak je podjela skupa podataka na dijelove na temelju aktivnosti kategoriziranih u prethodnoj fazi. Hodanje i drugi lagani pokreti smatraju se laganom aktivnošću, dok se nepomično sjedenje i stajanje smatraju neaktivnim. Trčanje, trčanje ili vježbanje je srednja aktivnost. Imputacija se zatim još jednom provodi, ovaj put s višestrukom linearnom regresijom i neuronskim mrežama.

Iako su rezultati dosljedni u svim kategorijama, točnost se povećala za ukupno 10% do 17%. U konačnici, rezultati nude zadovoljavajuću razinu točnosti klasifikacije i obećavaju njihovu primjenu u imputiranju podataka senzora. Neki od problema uključuju nedosljednosti u izvedbi

između laboratorija i okruženja slobodnog života, varijacije u somatotipu i sportskim navikama, koje imaju značajan utjecaj na rezultate imputacije i klasifikacije.

Poglavlje 5 zaranja u kritične aspekte specifikacije ograničenja semantičkih podataka i validaciju podataka u kontekstu zdravstvene skrbi. Poglavlje počinje prikazom materijala i metoda koji su korišteni, uključujući detaljnu raspravu o procesu prikupljanja podataka. Baca svjetlo na zamršenost analiziranja, provjere i potvrđivanja zdravstvenih podataka, naglašavajući važnost tih procesa u osiguravanju točnosti i pouzdanosti informacija. Ključna komponenta 5. poglavlja je istraživanje studije slučaja korištenja koja služi kao mehanizam verifikacije procesa. Rezultati ove studije detaljno su predstavljeni, usredotočujući se na specifične zdravstvene pokazatelje kao što su otkucaji srca, tjelesna temperatura i zasićenost kisikom. Ovo poglavlje pruža sveobuhvatnu analizu ishoda validacije, nudeći uvid u učinkovitost implementiranih procesa. Nadalje, Poglavlje 5 uvodi integraciju ovih validiranih podataka u postojeće zdravstvene informacijske sustave (HIS), pružajući pregled konceptualnog modela implementacije. Ovaj konceptualni okvir ocrtava besprijekornu asimilaciju validiranih zdravstvenih podataka, poboljšavajući ukupnu učinkovitost i pouzdanost upravljanja zdravstvenim informacijama. Personalizirani zdravstveni podaci, kao što su vitalni znakovi, prikupljeni kontinuirano i nenametljivo, potencijalno bi liječnicima mogli pružiti značajnu pomoć u praćenju ili postavljanju dijagnoze pacijenta. U ovom istraživanju korištena su dva skupa podataka:

- Skup podataka PMData dostupan je putem Simula Open Datasets,
- OxyBeat skup podataka prikupljen u svrhu ove studije kako bi se osigurao robusniji scenarij slučaja upotrebe dodavanjem brojnih dodatnih tipova podataka, s naglaskom na tipove podataka relevantne za COVID (tjelesna temperatura i zasićenost kisikom).

Oba skupa podataka prikupljena su pomoću uređaja za praćenje aktivnosti Fitbit Versa, tj. uređaja opremljenih senzorima koji se koriste za praćenje metrike povezane s fitnessom i zdravljem. Iako postoje mnogi čimbenici koji mogu utjecati na kvalitetu podataka, dva skupa podataka koji su korišteni prikupljeni su istim modelom nosivog uređaja i stoga su u velikoj mjeri usporedivi. Kopiranje podataka, identificiranje oštećenih podataka, tretiranje oštećenih podataka kao podataka koji nedostaju, imputiranje svih podataka koji nedostaju, a zatim izrada novog skupa podataka, sve su to koraci u procesu čišćenja podataka. U ovom slučaju, podatkovne točke s pouzdanošću jednakom 0 bile su one odabrane za imputaciju. Kao rezultat toga, pročišćeni podaci uključuju imputirane vrijednosti za podatkovne točke s razinom pouzdanosti nula, kao i izvorne vrijednosti za one s razinama pouzdanosti 1-3. Zatim se pomoću

provjere temeljene na Schematronu čisti podaci analiziraju i ispituju. Schematron je strukturna shema, jezik provjere valjanosti temeljen na pravilima i izražen u Extensible Markup Language (XML). Često se koristi za donošenje tvrdnji o prisutnosti ili odsutnosti specifičnih uzoraka u XML stablima i ima mogućnost stvaranja ograničenja na način na koji drugi jezici XML sheme, kao što su XML shema i definicija tipa dokumenta (eng. Document Type Definition, DTD), ne mogu. Provjera valjanosti sheme bila je jedina metoda provjere valjanosti koja se prvobitno koristila za XML dokumente. Ovo znači da se XML dokument smatra valjanim ako je zadovoljio provjeru valjanosti sheme. Iako osigurava ispravnu strukturu dokumenta, provjera valjanosti sheme ne može provjeriti uvjete i kriterije integriteta. Predloženi postupak za ovjeru osobnih zdravstvenih dokumenata provjeravao bi strukturu dokumenta prije ovjere sadržaja i karakteristika dokumenta. Konačno, potrebno je potvrditi sva dodatna ograničenja koja bi mogla dodatno postojati. Razvijena je nova generacija okvira standarda za razmjenu podataka iz elektroničkih zdravstvenih zapisa (EHR), nazvana specifikacija HL7 Fast Healthcare Interoperability Resources (FHIR). Ne postoji niti jedna globalna ontologija koja bi olakšala prijenos zdravstvenih podataka bilo koje vrste u široko korišteni standard HL7 FHIR, omogućujući interoperabilnost i poboljšavajući kvalitetu skrbi za pacijente i istraživanja, unatoč činjenici da su medicinski standardi razvijeni i prihvaćeni globalno (npr. , HL7 FHIR). Sheme će se stoga koristiti za specificiranje strukture zdravstvenih podataka i jamčenje standardizacije, što je bitno za uključivanje u službene EHR sustave. Potrebna je provjera i validacija podataka prije nego što se oni mogu uključiti u EHR. Proces je korišten za modeliranje procesa provjere valjanosti koji je osiguravao pridržavanje podataka propisima i standardima. To je postignuto:

- Definiranjem semantičkih ograničenja za tipove zdravstvenih podataka kako bi se zajamčilo pridržavanje standarda i propisa čineći informacije valjanima i relevantnima s medicinskog stajališta
- Definiranjem i modeliranje procedure za validaciju dobivenih podataka kako bi se mogli lako prenijeti i uključiti u službeni EHR. Naposljetku, ova je metodologija testirana u studiji slučaja korištenja uz korištenje postojećeg skupa podataka koji je sadržavao različite relevantne tipove podataka.

Dodatno, predložen je konceptualni pregled modela implementacije za integraciju u postojeći HIS. Za potrebe korištenja osobnih zdravstvenih podataka u individualiziranoj i preventivnoj zdravstvenoj zaštiti ključno je poštivanje propisa i standarda. Korištenjem postupka opisanog u ovom radu, novoobrađeni podaci mogu se službeno dodati u EHR u skladu s IHE standardima i propisima za relevantne vrste podataka. Model je u skladu s vodećim industrijskim

standardom, usmjeren je na integraciju podataka u EHR, nudi modul za čišćenje podataka i automatsku provjeru temeljenu na Schematronu.
Kako bi se zajamčila usklađenost sa standardima i propisima i kako bi informacije bile medicinski korisne i valjane, razvijena su semantička ograničenja za tipove zdravstvenih podataka. Predlaže se korištenje semantičke provjere valjanosti i validacijske procedure temeljene na Schematronu. Podaci će se moći prenijeti i uključiti u službeni EHR zahvaljujući metodi provjere valjanosti navedenoj u ovom istraživanju. Metoda je naknadno potvrđena pomoću skupova podataka koji uključuju nekoliko vrsta podataka povezanih sa zdravljem. Konačno, ove su komponente integrirane kao moduli u model središnjeg zdravstvenog sustava koji bi omogućio upotrebu IoMT podataka u EHR-u. Za korištenje osobnih zdravstvenih podataka za prilagođeno i preventivno liječenje neophodna je usklađenost sa standardima. Korištenjem postupka opisanog u ovom radu, proizvedeni podaci mogu se dodati formalnom EHR-u uz pridržavanje najnovijih IHE standarda za navedene vrste podataka. Model je u skladu s vrhunskim industrijskim standardom, fokusiran je na integraciju podataka EHR-a, nudi modul za čišćenje podataka i nudi automatsku provjeru valjanosti temeljenu na Schematronu. Time je zaokružen koncept arhitekture sustava za integraciju podataka s nosivih pametnih uređaja unutar središnjeg zdravstvenog informacijskog sustava, uključujući usklađivanje i validaciju podataka sa svjetskim standardima i propisima koji se odnose na EHR.

Poglavlje 6 posvećeno je ključnim temama sigurnosti i privatnosti u kontekstu zdravstvenih podataka. Poglavlje počinje istraživanjem različitih prijetnji sigurnosti i privatnosti koje su relevantne za zdravstvene informacijske sustave. To uključuje detaljno ispitivanje prijetnji povezanih s podacima pohranjenim na uređajima za praćenje i mobilnim uređajima, kao i potencijalne ranjivosti povezane s podacima koji se prenose bežičnim mrežama. Rasprava se proteže na regulatorni krajolik, baveći se propisima o podacima i standardima koji igraju ključnu ulogu u osiguravanju sigurnosti i privatnosti zdravstvenih informacija. Poglavlje 6 također razmatra izazove i razmatranja vezana uz pristup podacima s poslužitelja, naglašavajući potrebu za snažnim sigurnosnim mjerama za zaštitu osjetljivih zdravstvenih podataka. U biti, poglavlje pruža sveobuhvatan pregled pitanja sigurnosti i privatnosti koja su svojstvena upravljanju zdravstvenim podacima, nudeći uvid u potencijalne prijetnje i regulatorne okvire koji usmjeravaju zaštitu osjetljivih zdravstvenih informacija.

Cilj je bio omogućiti integraciju podataka prikupljenih s IoMT uređaja u elektronički zdravstveni karton (EHR) definiranjem modela sustava čišćenja i obrade podataka koji jamči

ukupnu kvalitetu i nepovredivost podataka. Transformacija podataka uključuje proces zamjene podataka koji nedostaju ili podataka koji su okarakterizirani kao nepouzdani procijenjenim podacima, pri čemu se vrijednosti koje nedostaju ili nepouzdane vrijednosti izračunavaju pomoću statističkih metoda i metoda strojnog učenja. Postojeći modeli transformacije podataka su analizirani, te je definiran poboljšani model kombiniranjem klasifikacijskih i regresijskih algoritama kako bi se osigurala točnost i preciznost podataka. Također, određena su semantička ograničenja, a podaci su potvrđeni putem shematrona kako bi se osiguralo da je put digitalnih podataka u skladu s međunarodnim normama i propisima za EHR. Hipoteza istraživanja bila je da odgovarajući proces čišćenja i transformacije podataka može osigurati točnost i kvalitetu podataka prikupljenih s IoMT uređaja te da se transformirani podaci mogu uskladiti s međunarodnim normama i propisima, naime Integrating the Healthcare Enterprise (IHE) i Health Level 7 (HL7); i uključena u EHR.

Izvorni znanstveni doprinos predloženog istraživanja je:

- Model sustava za čišćenje i transformaciju podataka koji se temelji na kombinaciji odabranih klasifikacijskih i regresijskih modela i njihovoj optimizaciji čime se osigurava točnost prikupljenih podataka.
- Specifikacija ograničenja semantičkih podataka i validacija podataka prikupljenih nosivim pametnim uređajima na temelju shematrona koji je u skladu s međunarodnim EHR normama i propisima definiranim Integrating the Healthcare Enterprise (IHE), kao i zdravstvenom razinom 7 (HL7).
- Prijedlog modela sustava za integraciju podataka s nosivih pametnih uređaja u središnji zdravstveni informacijski sustav koji uključuje usklađivanje i validaciju podataka s međunarodnim standardima i propisima vezanim uz EHR.

Ključne riječi: Internet medicinskih stvari, elektronički zdravstveni kartoni, senzori, standardizacija

Table of Contents

1. Introduction

Wireless Sensor Networks (WSN) and smart devices have been developing rapidly. Growing use of smart wearables and home automation devices has created numerous opportunities for developing novel high-quality solutions in many sectors, including eHealth. Research has begun on incorporating personalized health data collected through Internet of Medical Things (IoMT) devices [1][2], such as smartwatches or fitness bracelets, into an EHR which is a component of the central health information systems of many EU countries.

Internet of Things (IoT) devices' cutting-edge monitoring capabilities, which include continuous tracking of individuals' vitals or Activities of Daily Living (ADLs), can have a favorable impact on the quality of medical care they receive. Wearable electronics with embedded sensors are required to continuously and in real time monitor a person's vital signs.

Data can be collected from various sources, including different sensor types, as well as various manufacturers, each implementing their own proprietary data handling algorithms. The first task is to resolve any disparities, harmonize data, combine various data sets and merge them into a single, coherent aggregate. After the data have been aggregated, it must be ensured that all of it conforms to the same standard, regardless of its source. Lastly, an adequate data protection guarantee is essential given the sensitive nature of such data. One of the major challenges in making eHealth solutions impactful and valuable are gaining the trust of patients, protecting their privacy, and understanding the ramifications of possible self-diagnosing without the involvement of a professional.

1.1. Motivation

Sensors are a feasible and convenient way to unobtrusively and continuously monitor one's vital signs. Thus, use of sensors in the medical field can potentially provide significant assistance in finding the cause of a disease or establishing a diagnosis in complex cases.

- The potential of using user-collected health data is enormous, making it a paradigm that holds promise for improving people's health and wellbeing. Wearable sensors are omnipresent as they can be integrated into a wristwatch, an armband, adhesive bandaging tape or clothes. These sensors collect personal health data, such as heart rate, blood pressure, physical activity and body motion, temperature, respiration rate and oxygen saturation.

- The volume and diversity of health data are experiencing exponential growth with the adoption of electronic health records, medical wearable devices, and personal trackers; consequently, utilizing this data to its full potential would give us better understanding and facilitate optimal clinical decisions. This information would be extremely helpful to health practitioners since it would continuously provide them with vital, in-depth insight into patients' status, enabling an improved, more individualized approach to patient care.
- Furthermore, such a perspective change to a more proactive, rather than reactive medical service can lead to an overall decline in public health costs. The aim of the study was to investigate wireless sensors in the Internet of Things context with the goal to present a solution model for their use in a public healthcare system.

1.2. Objective and hypotheses of the research

The aim is to enable the integration of data collected from IoMT devices into the Electronic Health Record (EHR) by defining a model of cleaning and data processing system that guarantees the overall quality and inviolability of the data. Data transformation involves the process of replacing missing or data characterized as unreliable with estimated data, where missing or unreliable values are calculated by statistical and machine learning methods. Existing data transformation models have been analyzed, and an improved model has been defined by combining classification and regression algorithms to ensure data accuracy and precision. Also, semantic constraints have been specified, and data has been validated via a schematron to ensure that the digital data path is in line with international norms and regulations for EHR.

The research hypothesis is that the appropriate process of data cleaning and transformation can ensure the accuracy and quality of data collected from IoMT devices and that the transformed data can be harmonized with international norms and regulations, namely Integrating the Healthcare Enterprise (IHE) and Health Level 7 (HL7); and included in EHR.

1.3. Scientific contributions

The area of research is related to the issue of including personal health data which have been collected by wearable smart devices, such as personal tracking devices, into the EHR that today constitutes part of the central health information system of most European countries. In order for the data collected to be used in the EHR, it is necessary to guarantee their quality and inviolability. Therefore, data cleaning and processing is necessary. As part of the scientific work, a comparison of different models of health data cleaning has been made. A method for optimizing selected models of data cleaning and data transformation with the aim of improving accuracy and precision has been proposed. Furthermore, data compliance with existing standards and regulations must be ensured, which has been achieved by defining semantic constraints and validation via a schematron. A schematron is a block diagram, a rules-based validation language, expressed in XML.

Finally, these components are to be integrated as modules into a central health system model that would allow the use of IoMT data in EHR.

The original scientific contribution of the proposed research is:

- System model for data cleaning and transformation based on combination of selected classification and regression models and their optimization which ensures accuracy of data collected.
- Specification of semantic data constraints and validation of data collected by wearable smart devices based on a schematron that complies with international EHR norms and regulations defined by Integrating the Healthcare Enterprise (IHE), as well as Health Level 7 (HL7).
- Proposal of a system model for the integration of data from wearable smart devices within the central health information system, which involves harmonization and validation of data with international standards and regulations related to the EHR.

The following chapter gives an overview of Electronic Health Record; its definition, history, application and future potential, as well as applicable standards and regulations. Overview of Croatian implementation of EHR is also given. Third chapter provides relevant information regarding sensors and its use in healthcare, and wearable activity trackers specifically. Fourth chapter describes a system model for data cleaning and transformation, devised by comparing and optimizing data-driven models with use-case on ECG signals. Fifth chapter presents

specification of semantic data constraints and validation of data using the described process for parsing, verifying, and validating the data on the dataset collected for this purpose by a wearable activity tracker. Security and privacy concerns are explained in the sixth chapter. Finally, conclusions are given.

2. Electronic Health Record (EHR)

2.1. Definition

Information Technology has become ubiquitous in healthcare, as many healthcare providers switch from paper medical charts and documents to IT solutions in order to simplify administration, improve management of medical records and, ultimately, optimize the provided medical care [24]. The need for improved and more efficient healthcare is increasing, as a result of increase and aging of the population, pandemics, and complex health issues. Thus, Electronic Health Record (EHR) would provide complete patient's information which would be readily available across multiple healthcare organizations.

EHR is defined by ISO Technical Report (ISO/TR) 20514 as “A repository of information regarding the health of a subject of care in computer processable form, stored and transmitted securely, and accessible by multiple authorized users. It has a standardized information model, which is independent of EHR systems. Its primary purpose is the support of continuing efficient and quality integrated healthcare and it contains information, which is retrospective, concurrent, and prospective”.

EHR contains, patient’s medical history, including previous medical issues, encounters, and treatments, allergies, immunizations, vital signs, laboratory reports and radiology and diagnostic imaging, all in a digital form. As its purpose is to provide care to a patient across multiple healthcare organizations, it is not owned by any single entity. Hence, it must be capable of integrating data from mutually independent healthcare systems and allow interoperability among them.

Electronic Medical Record (EMR) is usually used interchangeably with EHR, although it is important to note that it is sometimes considered as a "snapshot" of a larger EHR, e.g., record related to a single encounter.

The advantages of using an Electronic Health Record are numerous. It improves the efficiency of the medical processes and workflow, especially if multiple healthcare providers are involved in the treatment (Figure 1). This means better patient care and improves quality of care, as less time is used on reporting and charting data. Paperless system requires less physical space, and overall reduces healthcare delivery costs. It also provides large amounts of data that can be used for disease analysis, treatment monitoring and preventive measures on a wide scale. Through

its security control mechanisms; authorization and audit, it offers superior security to paper charts.

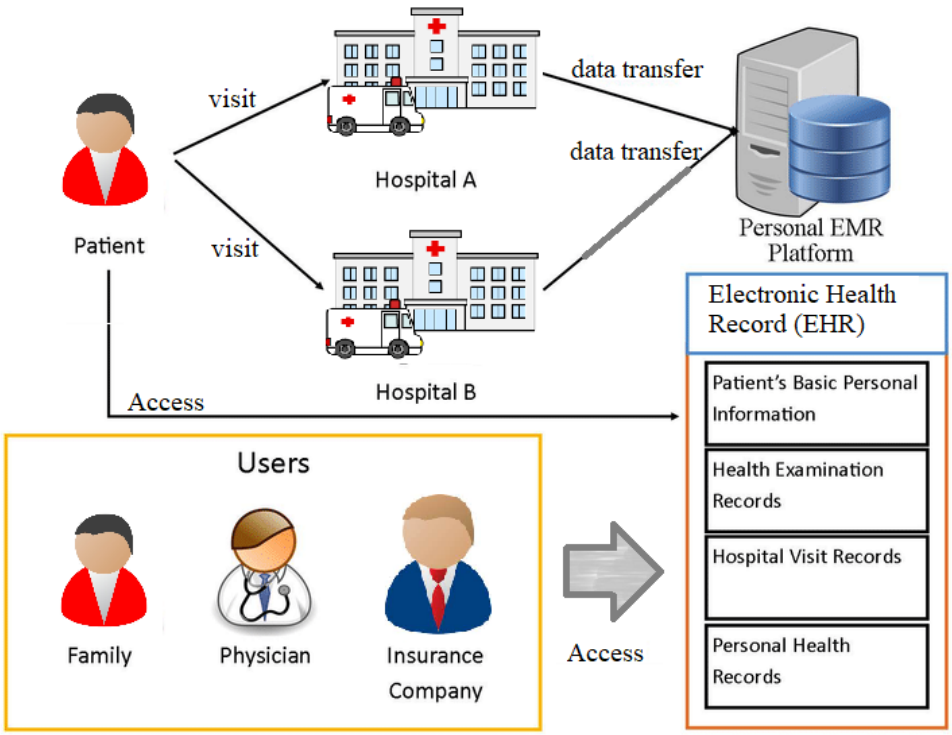

Figure 1. Multiple healthcare providers using EHR

Personal Health Record (PHR), however, is a health record, i.e., a collection of health-related information, which is controlled and maintained by the patient. Such record may be generated from various sources, e.g., EHR, physicians, the patient, or third-party applications. The following chapter gives an overview of PHR, its differences compared to the EHR and the potential of interoperability between the two.

2.2. Personal Health Record (PHR)

The ISO TR14639-2:2014 defines PHR as the "representation of information regarding or relevant to the health, including wellness, development, and welfare, of a subject of care, which may be stand-alone or integrating health information from multiple sources [25]. Loosely, PHR can mean any patient-controlled health information, no matter the form, structure or where it is stored. The term PHR has first been mentioned in 1978 [25], but a digital PHR has had its beginnings much later [26]. In 1994, P. Szolovits, J. Doyle, and W. J. Long (Laboratory for

Computer Science, MIT), I. Kohane (Boston Children's Hospital), and S. G. Pauker (New England Medical Center) proposed the Personal Internetworked Notary and Guardian (PING) [27], a free, open-source Personally Controlled Health Record (PCHR). PING was controlled and maintained by the patient, who is responsible for storing the encrypted information on whichever site they chose. Accessing the record, authentication, and authorizing the users and the process of encryption was to be done by publicly accessible PING servers. The main objective was to pass the control over to the patients themselves. Notable attempts at launching a widespread PHR include Google Health and Microsoft HealthVault PHR.

Google Health was launched in 2008 but discontinued in 2011 [28]. Google Health lacked the ability to import data from third-party sources and was not directly connected to the health services providers' network. Users would, often manually, enter health-related information such as medical conditions, allergies, and medications.

Subsequently, Microsoft launched HealthVault PHR in 2007, but discontinued it in 2019 [29]. HealthVault PHR gave the patients the ability to control and authorize access to their PHR record. In 2014, Microsoft Band (Figure 2), a fitness band, was launched. Microsoft Band enabled aggregation and integration of data into the HealthVault via MyFitnessPal smartphone application which tracked a person's diet and exercise. It was, however, discontinued in 2016. In 2018, the application suffered a security data breach of 150 million accounts. HealthVault did support some early pre-FHIR standards, i.e., the Continuity of Care Document (CCD) and the Continuity of Care Record (CCR).

Finally, in 2018, Apple added a "Health Records" section within the "Health" iOS application where users could see their medical data, such as allergies and conditions, from multiple health service providers within the USA, on their smartphones.

Figure 2. Microsoft Band fitness band

Since clinical institutions supply the majority of the medical data, patients have little control over or access to their own medical records. Patients or caregivers may need to maintain track of medical information that isn't recorded in the EHR, such as measurements or observed symptoms. As a result, monitoring, diagnosis, and treatment would all benefit from an effective and efficient Personal Health Record System (PHRS) that enables patients or caretakers to continuously monitor and manage the personal health record. Additionally, statistical anonymized data improves the overall healthcare system. Thus, personalized medical services, as well as healthcare as a whole would benefit from interoperability between these domains (Figure 3).

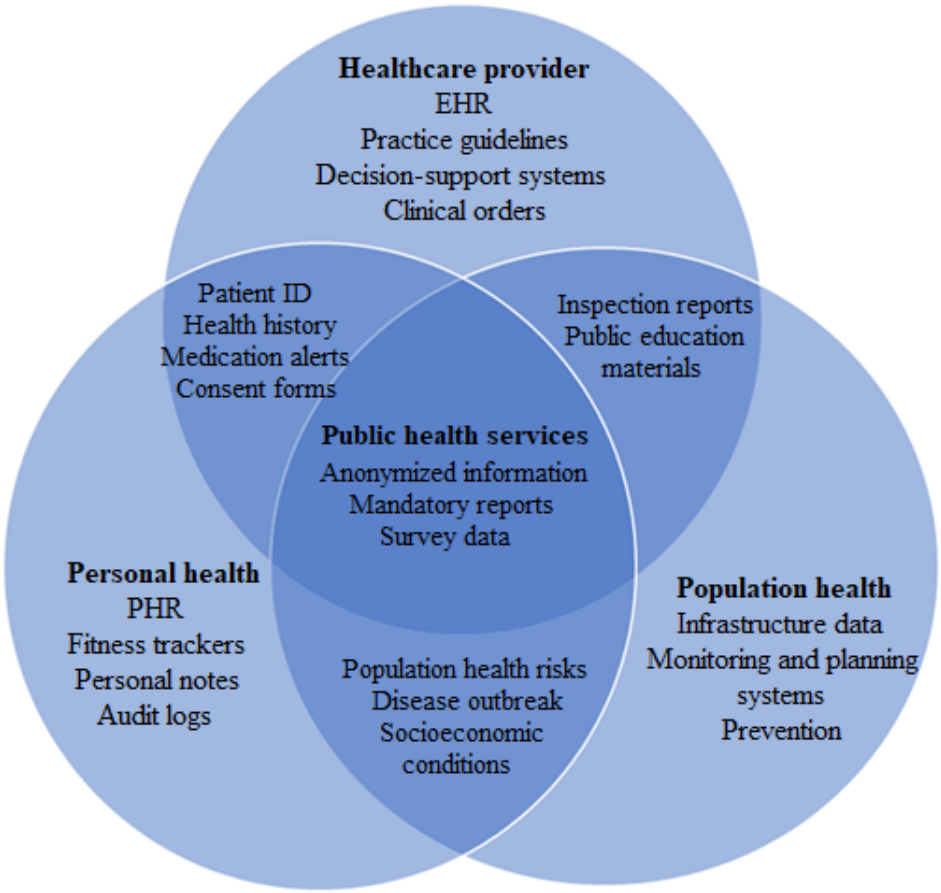

Figure 3. Overlapping of health domains

2.3. History

Traditionally, as of 1900-1920, health records were notes manually written on paper and organized into sections. New technologies developing in the 1960s and 1970s allowed for the emergence of what is today known as Electronic Health Record (EHR). Originally, EHRs were created and used at academic medical facilities. The first notable attempt to streamline and improve the patients' medical files was the Problem-Oriented Medical Record (POMR) created by Dr. Lawrence Weed in 1968 [31]. POMR consisted of:

- Database,
- Complete problem list,
- Initial planning,
- Notes of daily progress,
- Discharge summary.

During 1970s, the first EHR was developed at The Regenstrief Institute in Indianapolis after consulting with computer science experts from Purdue University [32] and integrated:

- Prescriptions and medication orders,
- Procedures done,
- Nursing orders,
- Diets,
- Laboratory tests,
- X-ray scans.

Development of EHRs between 1972 and 1992 included hierarchical or relational databases and where EHRs were deployed on large mainframe computers with limited storage [33]. This meant additional storage was necessary, either through disks or tape, as well as dedicated wired terminals. Very few early EHRs supported prescriptions and notes but were instead focused on laboratory results.

Web-based EHRs started to appear between 1980 and 1990, as a consequence of hardware becoming more affordable, powerful, compact and the appearance of the Internet, facilitated quicker and more convenient access to information [34].

As third-party applications were beginning to be used within EHRs, standards were required. In the early 2000s, Health Level Seven (HL7) and IEEE P1157 MEDIX served as the primary interface standards. This was necessary to disambiguate data element definitions and use standardized dictionary codes. Eventually, the HL7 standard was updated and expanded to

incorporate numerous systems, such as laboratory results or electrocardiogram (ECG) [35] and included protections:

- Automatically backup the data,
- Automatic log off,
- Data encryption,
- Audit logs,
- Access control.

Between 1991 and 2005, large healthcare providers, academic researchers, and government organizations started to push for the use of EHR, first in the USA and, to lesser extent, Canada, followed by the United Kingdom, Switzerland, the Netherlands and Norway [36]. Since 2005, EHR use has been steadily increasing in the majority of European countries, Australia, and Asia.

Most modern EHRs are web-based with client and server side, use relational databases, provide secure authorization-based data access, and allow interoperation of multiple entities, such as hospitals, physicians, or pharmacies. Figure 4 shows the timeline of EHR development.

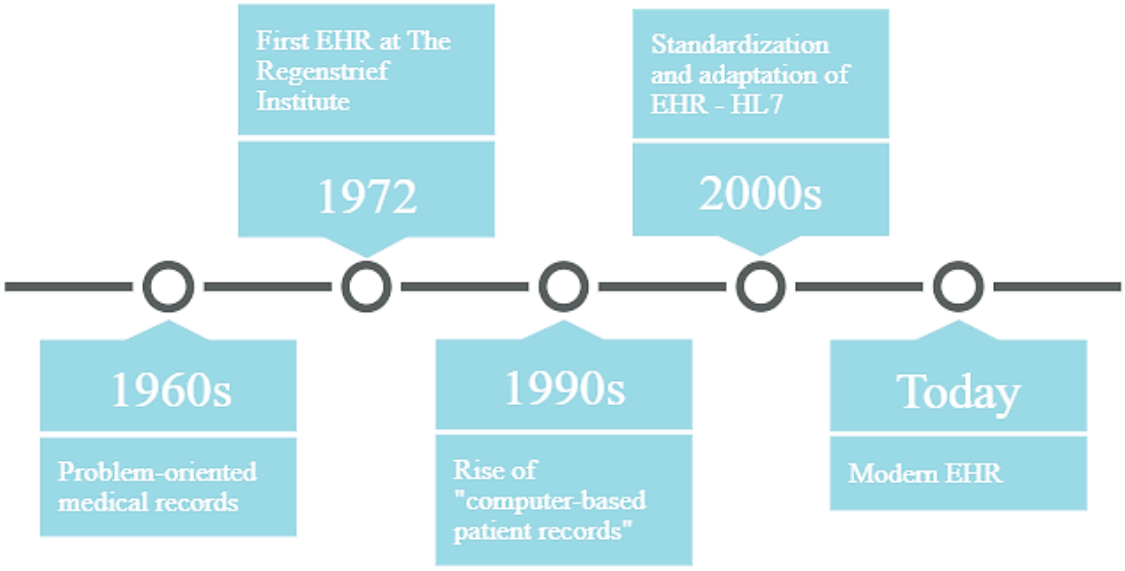

Figure 4. EHR timeline

2.4. Applications and future potential of EHR

The most significant feature of the smart wearables embedded with sensors is undoubtedly the capability to track health. An increase in the demand for wireless fitness wearables is driving the market, its cause being consumers' increased health awareness. With an emphasis on vital signs and fitness tracking activity, wearable sensors represent a significant component of wearable technology. A fitness or activity tracker is a device worn on the human body, made to track and document health-related data and an individual's fitness activity, including heart rate, body temperature, steps taken, stairs climbed, calories burned, etc. The wearables usually offer wireless connection from the wearer's phone.

According to the report [3], the global market for wearable fitness trackers will expand by an estimated 78 million euros from 2020 to 2027, with a Compound Annual Growth Rate (CAGR) reaching 22.6%. Equivalently, IoT-capable smart watches are a developing trend because not only do they function as a stand-alone device but also enable users to communicate with other IoT-devices, significantly enhancing the quality of service. Every smartwatch has a fitness tracker built in, therefore providing its user with a variety of capabilities, including serving as a health-related data monitor. The report [4] estimates that the global smartwatch market had 43.87 million in shipping volumes in 2018 and is projected to increase up to 108.91 million by 2024, with a CAGR of 14.5% over the forecast period between 2019 and 2024. This demonstrates that using wirelessly collected personal health-related data in a central health information system shows promise since it may provide improved understanding of the patients' states, which would help the decision-makers provide optimal, more individualized treatment. Nonetheless, the fact that the system deals with such sensitive information presents a number of difficulties because it must be reliable, secure, and error-free when handling the data.

According to reports [5] and [6], the market for the IoMT is anticipated to develop at a 19.9% CAGR and reach 455.3 billion euros by 2025. IoMT encompasses all medical devices as well as the software that provide health- and healthcare-related services. Telemedicine, also known as remote healthcare, and monitoring an individual's activities or health status via wearable smart devices are examples of potential applications. Body-worn smart electronics equipped with sensors, e.g., smart watches, fitness bands, smart glasses or rings can help monitor health- and fitness-related data including calories burned, heartbeat rate, or glucose levels, among many other parameters. Additionally, a number of mobile platforms and applications have already been created in an effort to support Coronavirus Disease 2019 (COVID-19) impact management.

2.5. Impact during COVID-19

COVID-19 is a respiratory illness with most common symptoms being fatigue, fever, dry cough, dyspnea (shortness of breath). In more severe cases it can lead to Acute Respiratory Distress Syndrome (ARDS). It originated in Wuhan City, China in December 2019. Up to December 2021, according to the World Health Organization (WHO), 613,942,561 confirmed cases of COVID-19 were reported worldwide, 6,520,263 of those instances ending in death [7]. The Americas, South Asia, and Europe were the areas that were most impacted during the period, as shown in Figure 5.

Brazil, India, and the United States of America have recorded the largest number of instances (Figure 6). According to findings, human-to-human respiratory droplet transmission is the primary means by which the COVID-19 virus spreads. It quickly had a devastating effect on the world economy, leading several nations to impose travel restrictions and go into lockdown, which had a detrimental effect on the industrial sectors [8]. The following are the sectors that suffered the most:

- Automobile industry,
- Transportation systems, especially airplane transportations,
- Construction,
- Tourism,
- Public healthcare systems.

When attempting to provide services to an increasingly growing number of patients, public healthcare systems in particular encountered a shortage of resources. This put an unparalleled burden on the medical staff, which were also more likely to become infected with the virus. Personal Protective Equipment (PPE) and medical ventilator shortages plagued most hospitals. The goal was "flattening the curve," or to limit the spread of the disease, in order to allow medical professionals to maintain delivering the required treatment without feeling overburdened. In an attempt to control the pandemic, traditional community-containment methods were used. By limiting interpersonal contact, they hoped to contain epidemics and lower infection rates and numbers. These actions include:

- Social distancing - preventive measures such as discouraging needless interpersonal contact, regardless of a one's COVID-19 status because people without usual symptoms

may carry the virus. Enforcement of the principle of maintaining a safe distance among people in both open and enclosed public spaces,

- Quarantine - restriction of the mobility of people who did not test positive but that may have come into contact with a person that did,
- Isolation – which involves separating those who are confirmed to be COVID-19 positive from those who are not in order to stop the spread of the illness,
- Lockdown restrictions, sometimes known as a "stay at home" order, is a government directive that instructs citizens to stay at home as much as they can (not applicable to essential tasks). It can be regarded as a tactic of mass-quarantine used to control the pandemic.

In addition, many businesses and governments require that masks be worn while traveling and in enclosed or congested areas. Frequent hand washing, avoiding touching the face, and covering the mouth and nose with an elbow or tissue when coughing or sneezing are other COVID-19 hygiene precautions.

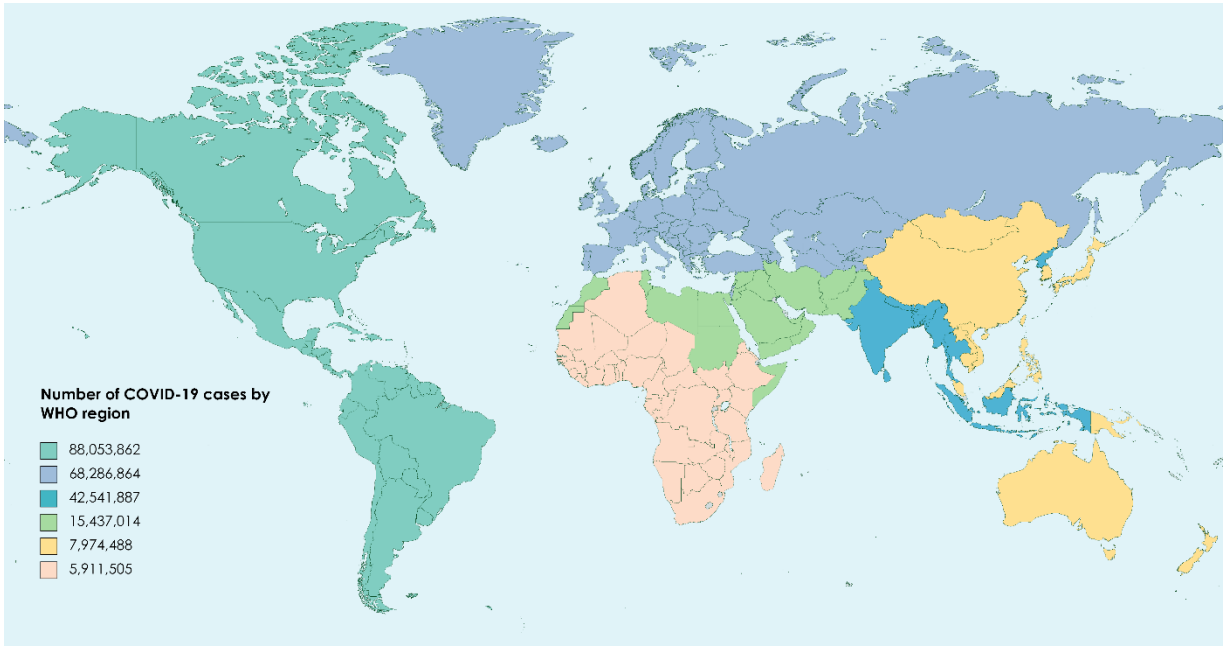

Figure 5. Number of confirmed coronavirus cases by WHO region in 2021

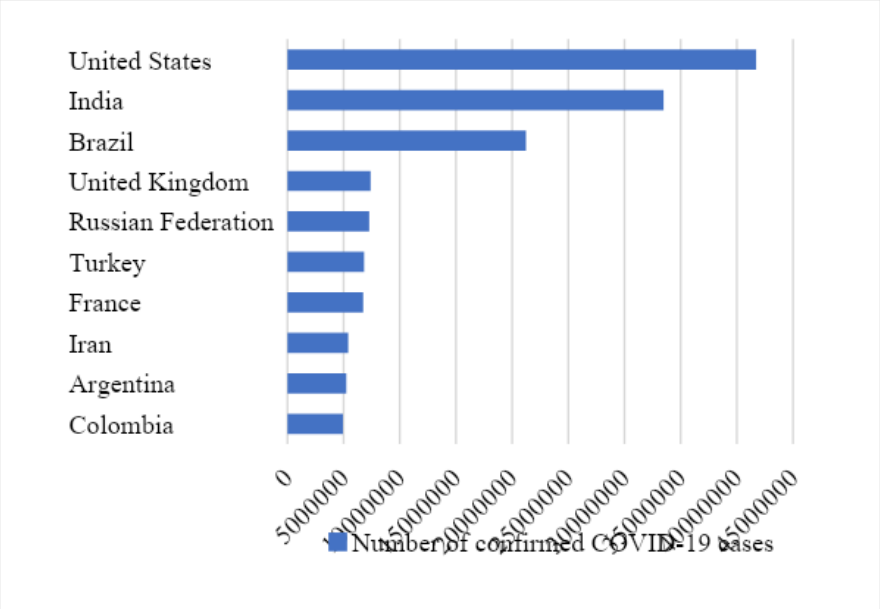

Figure 6. Top ten countries with most cases of coronavirus reported in 2021

Information technology, as WHO reports, plays a significant role in improving response of the health system to the pandemic. It helps with:

- Monitoring, including remotely, of patient's health,
- Telemedicine, i.e., provision of remote care,
- Social distancing,
- Contact tracing, i.e., identification of people who might have had contact with infected individuals.

Using Bluetooth, blockchain and spatial data analysis are some of the methods used to implement contact tracing solutions. A thorough overview of applications used to trace coronavirus contact is provided in [9], and Figure 7 depicts their process. The keys are just a randomized sequence of letters and numbers; both the device and the user are anonymized. To ensure the individual's anonymity throughout, they change multiple times an hour. Mobile smartphone devices exchange the keys using communication technology (e.g., Bluetooth) when a user is in direct proximity to another application user. As a result, the software can keep records of contact made without revealing the identity of the individuals in question.

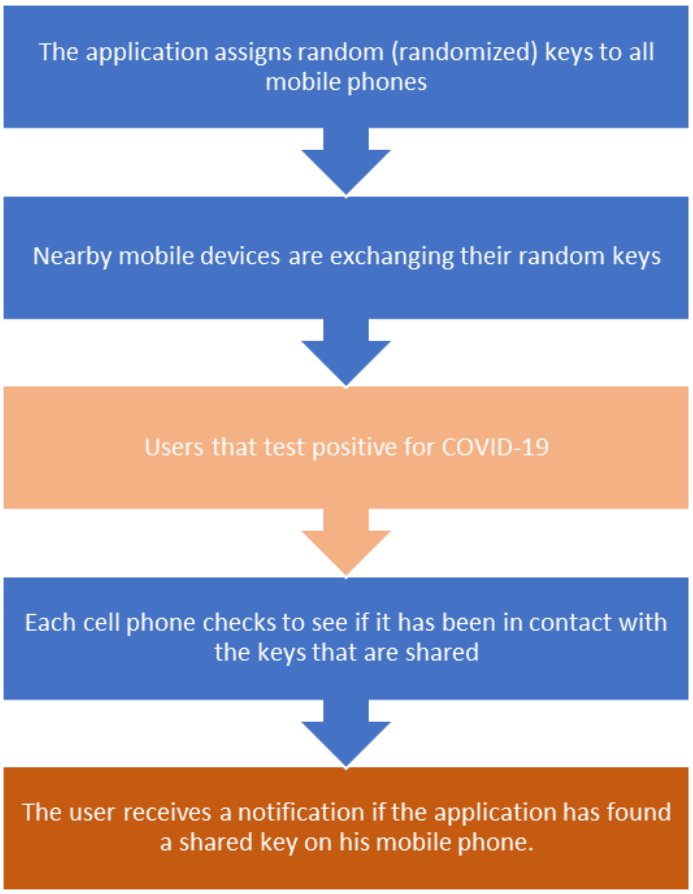

Figure 7. Contact tracing process

When a user tests positive for COVID-19, they can send their keys which were broadcasted previously to other users in proximity. Similarly, the code for an individual can be entered by a healthcare practitioner upon obtaining a positive laboratory result. The software on a regular basis compares the keys that were sent to the server following a positive test result and the keys kept on the user's smartphone device, collected from other users that the person was in contact with. The application determines if a person has been exposed to and is at risk of contracting the disease. Upon a positive match, i.e., someone has had contact with another individual who has tested positive and shared their keys, the former will be informed and given guidance on what to do. The privacy of the user is preserved as the application doesn't require identifying information. [10] leverages IoT to implement messaging systems for social distancing. Smart medical masks [11] use temperature and strain sensors and serve to detect the illness as soon as possible, and [12] has used commercially available particulate matter sensors to detect pathogen particles in the immediate vicinity of the user.

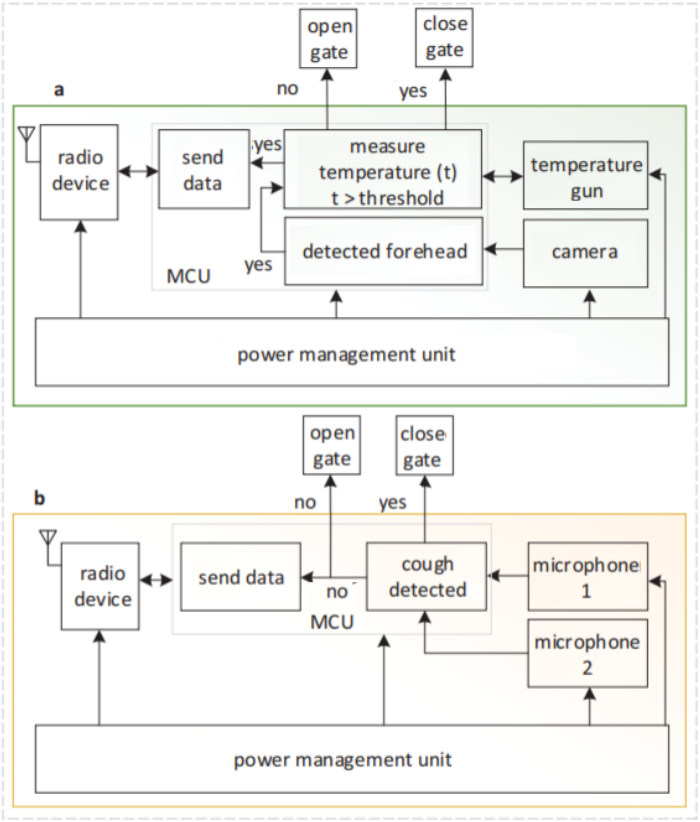

Figure 8. Schematic for Temperature Sensing Unit (TSU) and Cough Detecting Unit (CDU) [13]

The most common signs of COVID-19 are a dry cough and a raised body temperature. As a result, [13] proposes that each entrance point at transportation hubs including airports, railway stations, and bus terminals be equipped with two units, first one being the Temperature Sensing Unit (TSU), which uses a camera to identify an individual's forehead before measuring temperature using a temperature gun. Depending on whether the recorded temperature was below or over the threshold, the microcontroller then transmits the order to "open" or "shut" the gate. The Cough Detecting Unit (CDU) is the second device; it uses two microphones and a microcontroller to identify cough sounds. Figure 4 contains the schematics for each of the units. Similar to this, [14] suggests an inexpensive scanning device, consisting of several sensors, a camera, and a microcontroller, for COVID-19 detection that would combine Artificial Intelligence (AI) and IoT. Due to its dimensions and affordability, it would be able to scan a significant number of persons successfully.

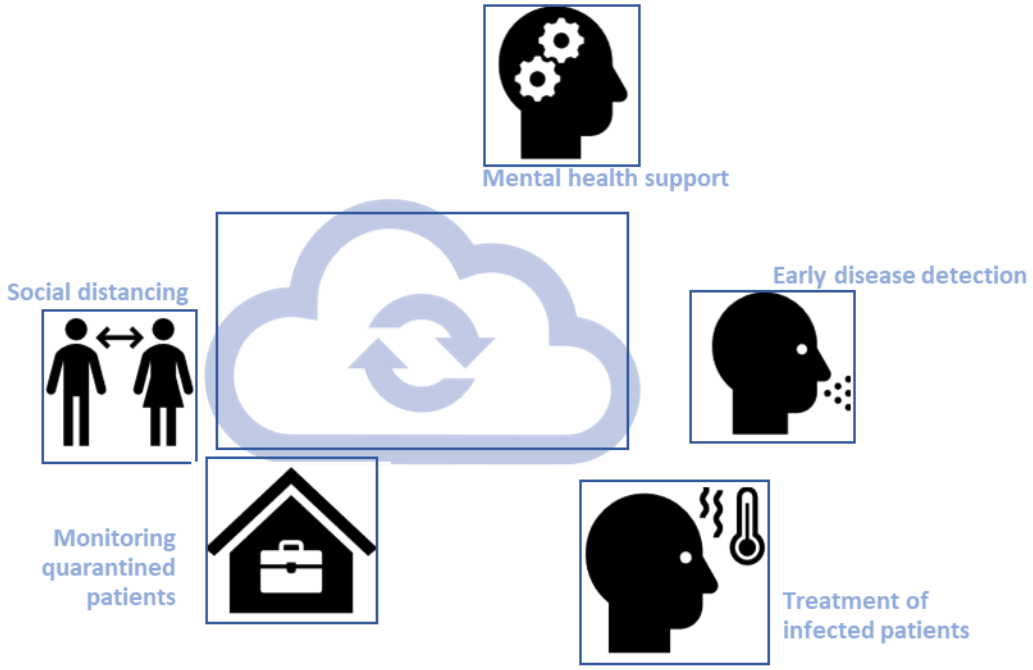

Figure 9. IoT application used in COVID-19 mitigation

Whereas certain approaches attempt to "flatten the curve" by implementing social distancing, contact tracing, or prompt detection, others provide assistance by remote monitoring of infected or home-confined individuals. [15] suggests a patient monitoring system that uses sensors which measure respiration and heartbeat rates, body temperature and oxygen saturation, to supervise possibly contagious, quarantined individuals. [16] integrates a smart wristwatch positioning and data collected by fitness trackers with a telemedicine platform to observe positive persons staying at home. In terms of mental health, machine learning algorithms have been utilized by [17] to provide mental health assistance and treatment recommendations. Figure 9 provides an illustration of these.

Finally, the fact that confidential personal information is being exchanged raises worries about security and privacy. The following are the top three [18] risks to data security, known as CIA:

- Confidentiality - data must be kept confidential and must not be shared with unauthorized parties,
- Integrity – tampering with data is not allowed,
- Availability - data has to be ready to be accessed whenever needed.

Additionally, in the context of location, security encompasses:

- Device - the physical device must be protected, for example using a password,
- Communication channel - data must be maintained secure throughout the process of transmission between the user's device and the cloud,
- Cloud – storing the data and maintaining it secure by encrypting it.

To guarantee secure transmission when sharing and maintaining pandemic data, the COVID-19 stochastic model was devised [18]. Furthermore, all software that collects personal information must be compliant with European Union's laws and guidelines, i.e., General Data Protection Regulation (GDPR), which means all data must be:

- provided voluntarily by the user,
- used transparently,
- kept temporarily until not needed,
- kept securely, e.g., encrypted,
- use only pseudo-anonymous data.

Finally, the software must request for the user's explicit consent, which may be revoked at any moment. The person shall be able to freely govern the data they wish to disclose, and the use of software must be entirely voluntary in all its phases.

Fifth Generation (5G), and future Sixth Generation (6G), mobile networks will have better performance, i.e., faster speeds and reduced latency, are more reliable and have increased availability. Table 1 gives a comparison of 4G, 5G and 6G. Due to faster speeds and lower latency, 5G and 6G network technology will provide an improved connection for IoMT devices-to-cloud communication and thus improved performance and QoS in the telemedicine. In conjunction with solutions from IoMT and AI, 5G and future 6G can make patient monitoring and virus tracking easier. Additionally, further studies can be performed using the acquired data. This is how China has already been utilizing commercial 5G network in initiatives, such as the COVID-19 internet platform for teleconsultation, diagnosis, and treatment [19]. Reconfigurable Intelligent Surfaces (RIS) for 6G are being examined in projects that remotely monitor health, e.g., human posture recognition as one of the many options that 6G provides for smart healthcare [20]. Additionally, edge intelligence, which applies AI to decentralized data and IoMT devices, can be used in analyzing COVID-19 data [21]. Lastly, mURLLC—massive Ultra-Reliable Low-Latency Communication—will improve communication between, wearable sensors and cloud servers [22]. Thus, it is likely 6G will impact healthcare in the future. In the following five years, a significant switch to 5G is anticipated. While this is happening, initial reports [23] anticipate that 6G will be deployed commercially in 2028 and 2029.

Table 1. Comparison of 4G, 5G and 6G networks

	4G	5G	6G
Peak data rate	1 Gbps	10 Gbps	1 Tbps
E2E latency	< 100 ms	< 10 ms	< 1 ms
Max spectral efficiency	15 (bps) Hz	30 (bps) Hz	100 (bps) Hz
Max frequency	6 Ghz	90 Ghz	10 THz
Mobility support	350 km/h	500 km/h	1000 km/h
Architecture	MIMO	Massive MIMO	Intelligent surface
Satellite integration	No	No	Yes
AI	No	Limited	Yes
Autonomous vehicles	No	Limited	Yes
Haptic communication	No	Limited	Yes

2.6. Standards and regulations

Multiple widely adopted health data standards exist [38]. Additionally, some healthcare providers use their own proprietary standards, which have little to no integration with others. Many countries recommend adoption of a specific, recognized standard, interoperability across different healthcare institutions being the key objective. Nonetheless, as many of them are incompatible with one another, implementing open and globally acknowledged standard does not ensure compatibility [39]. The patient's data can be challenging to integrate, despite continuous efforts to encourage the usage of the standards and the development of open specifications [40]. The complexity of integration of PHR data is greater when it combines data from a patient's wearable technology. This is due to the fact that PHR can contain data from many sources, irrespective of the healthcare provider. Most popular standards are HL7 standards (newest being HL7 FHIR) and OpenEHR.

2.6.1. HL7 FHIR

Health Level Seven (HL7) [41] is ANSI-accredited family of standards which is extensively used worldwide for data exchange in health and medicine IT systems. In 2003, the Health Level 7 EHR Special Interest Group (HL7 EHR SIG) developed the HL7 EHR-System Functional Model (HL7 EHR-S FM) which contains guidelines for functional expectations of an EHR system. It defined a functional profiles template which can be used by the users to add new functional profiles. Standards developed by HL7 were HL7 v2.x, HL7 v3, Clinical Document Architecture (CDA), Continuity of Care Document (CCD), and the latest, Fast Healthcare Interoperability Resources (FHIR).

The primary goal of HL7 v2.x standards was to enable electronic exchange of healthcare data among different healthcare entities, i.e., including data transfer onto removable storage media, such as disks or tape. HL7 v2.x standard simplified applications' interface implementation. Specifically, in HL7 v2.5 standard there are twelve different functions represented as message structures defined, as follows:

- Patient Administration (ADT), information related to admitting or discharging a patient, visits, queries, and changes in status,
- Order Entry, messages for requests, updates, or query orders, related to, e.g., medications, or laboratory tests, given by physicians. Specifically, orders can be general (ORM), query (OSQ), general clinical (OMG), laboratory (OML) and imaging (OMI),
- Financial Management includes adding or changing the billing account (BAR), viewing insurance payments and detailed financial transaction (DTF), and accounts management,
- Observation Reporting defines search, query, and report of laboratory test results (ORU and OUL) or clinical observations (QRY and ORF),
- Master Files notification (MFN) and query (MFQ) which preserve information and demand that any made update to such files needs to be announced to all entities involved in order to synchronize information,
- Medical Records Management (MDM) contains messages related to document management, whether appending, archiving, canceling, or authentication, and updating the status of documents, e.g., new, updated, restricted,
- Scheduling consists of messages for scheduling requests and response (SRM and SRR), and scheduling query and response (SQM and SQR) of any resource within healthcare facility, such as visit, laboratory test or imaging,
- Patient Referral contains messages for referring a patient across different medical facilities and transferring the relevant information, such as, patient information (RQI), clinical information (RQC), unsolicited insurance information (PIN), patient authorization (RQA) and patient referral request (REF),
- Patient Care supports exchange of information related to health problems and treatments with messages for patient goal (PGL), problem (PPR), pathway (PPP), and query patient care problem (QRY),
- Clinical Laboratory Automation interfaces medical equipment with the Laboratory Information System (LIS) and contains messages for updates of automated equipment status (ESU), equipment inventory (INU), and specimen (SSU),

- Application Management Query (NMQ) and Data (NMD) consists of messages related to application-level information exchange, e.g., software version or system clock,
- Personal Management (PMU) relates to administrative activities concerning staff, such as permission changes, i.e., adding, updating, deleting, deactivating and terminating personal records.

HL7 v2.x defines two basic message types:

- Event means real event occurring which involves transfer of health information. The response is an acknowledgment message (ACK). For example, “Register a patient on a clinical trial“ event is defined by CRM_C01 where CRM is the message type and C01 is the event code.
- Query queries healthcare data and the response contains the requested information. For example, "Query for Master File Record - Test/Observation" is represented by MFQ_M03 and the respective response message is MFR_M03.

Table 2. shows real-world scenarios of various events and respective use of HL7 v2.x messages generated.

Table 2. HL7 use-case scenario

Event	**HL7 message**
Person visits the hospital.	Registration of the patient (send Patient Register event).
Doctor diagnoses fever and cough, orders COVID-19 test.	Send Order message to laboratory system.
Receive results and report. Diagnose the patient with COVID-19.	Send a Patient Admit message.
Prescribe medications and treatment.	Send a Medication Order message to the pharmacy system.
Discharge the patient after recovery	Send a discharge message.

However, HL7 v2.x doesn't cover security, confidentiality, nor data integrity of health-related information exchanged. It does not specify any security mechanism or encryption algorithms. Applications must support their own security and privacy mechanisms. Furthermore, HL7 v2.x does not support anonymization of data. Due to these issues, another standard was made.

HL7 v3 has domains similar to HL7 v2.x functions, however it does introduce message wrappers which provide more information and describe the role of the sender or receiver and its messages XML-based. HL7 v3 introduces Reference Information Model (RIM) which is object-oriented and consists of:

- Classes, class attributes, and class relationships,
- State transition models, and
- Data types and constraints.

Aside from the definition of a common message specification, HL7 v3 developed is Role Based Access Control (RBAC) specification for role-based access and permissions which provides security. HL7 v3 also includes an implementation guide for implementation on a Service-Oriented Architecture (SOA).

FHIR expands on previous standards by employing RESTful web services and open web technologies and including JSON and RDF data formats alongside previously used XML. FHIR is a standard for healthcare data exchange; for exchanging healthcare information electronically. Example is given below for HL7 FHIR for fetching administrative data in the form of a sequence diagram (Figure 10) [41].

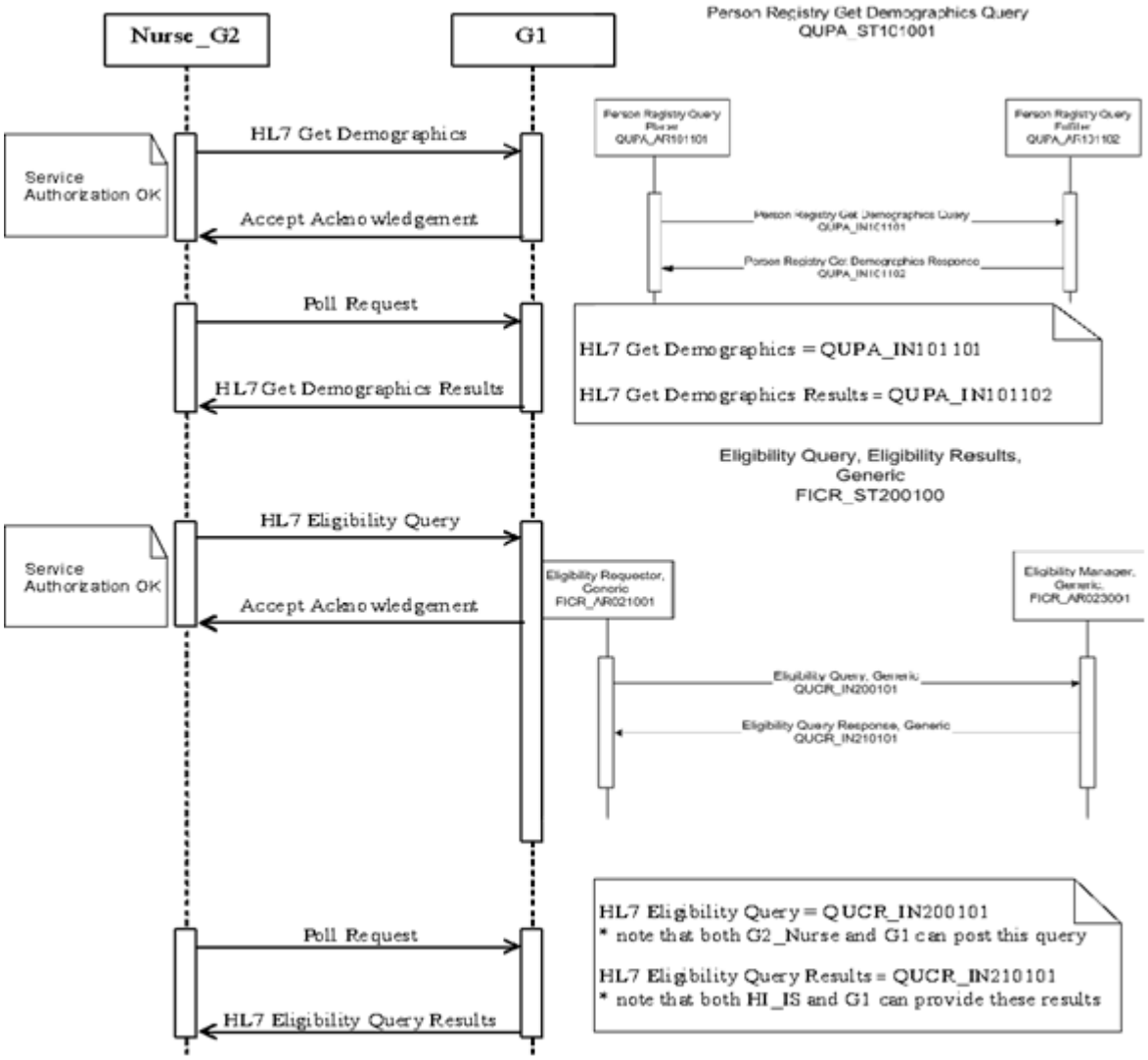

Figure 10. Sequence diagram for fetching Patient Administrative data [41]

The basic building block in FHIR is a Resource, it consists of a set of metadata, standard data and a human readable part (as shown in Figure 11) [41].

Figure 11. HL7 FHIR Resource [41]

A new generation of standards framework for the interchange of data from Electronic Health Records (EHRs) was developed, called the HL7 Fast Healthcare Interoperability Resources (FHIR) specification. There isn't a single global ontology that would facilitate transferring healthcare data of any type into the widely used HL7 FHIR standard, enabling interoperability and enhancing the quality of patient care and research, despite the fact that medical standards are developed and adopted globally (e.g., HL7 FHIR). Schemas will therefore be used to specify the structure of health data and to guarantee standardization, which is essential for inclusion into official EHR systems. Verification and validation of the data are required before it can be incorporated into the EHR.

Eventually, the release of HL7 Fast Healthcare Interoperability Resources (FHIR) offered the first normative for:

- the RESTful API, the XML and JSON formats, as the basic data types,
- the Terminology Layer (Code System and Value Set),
- the Conformance Framework (Structure Definition and Capability Statement), and
- the key resources – Patient and Observation.

FHIR is developed under the 80/20 rule, which means that existing resources cover 80 percent of all possible needs and the 20 percent that remain are too specific use cases which will be handled with FHIR extensions.

HL7 standards are the backbone of the National Health Information Service (HIS) in many countries, including Croatia, Finland, Norway, Sweden, Iceland [196], and India [197].

2.6.2. openEHR

openEHR [55] is a health data open standard that outlines how to manage, store, retrieve, and exchange health data EHRs. All of a patient's health information is kept in a patient centered EHR. In contrast to HL7, the openEHR specification is not focused on the sharing of data across EHR systems, with the exception of the EHR Extract specification [54]. The openEHR specification [55] covers clinical workflow, demographics, information, and service EHR models, and archetypes. It is intended to serve as the cornerstone of a distributed, versioned EHR system that is medically and legally sound. The openEHR standard consists of (Figure 12) [56]:

- information models (reference model),
- the archetype formalisms,
- the Archetype Query Language (AQL),
- service models and APIs.

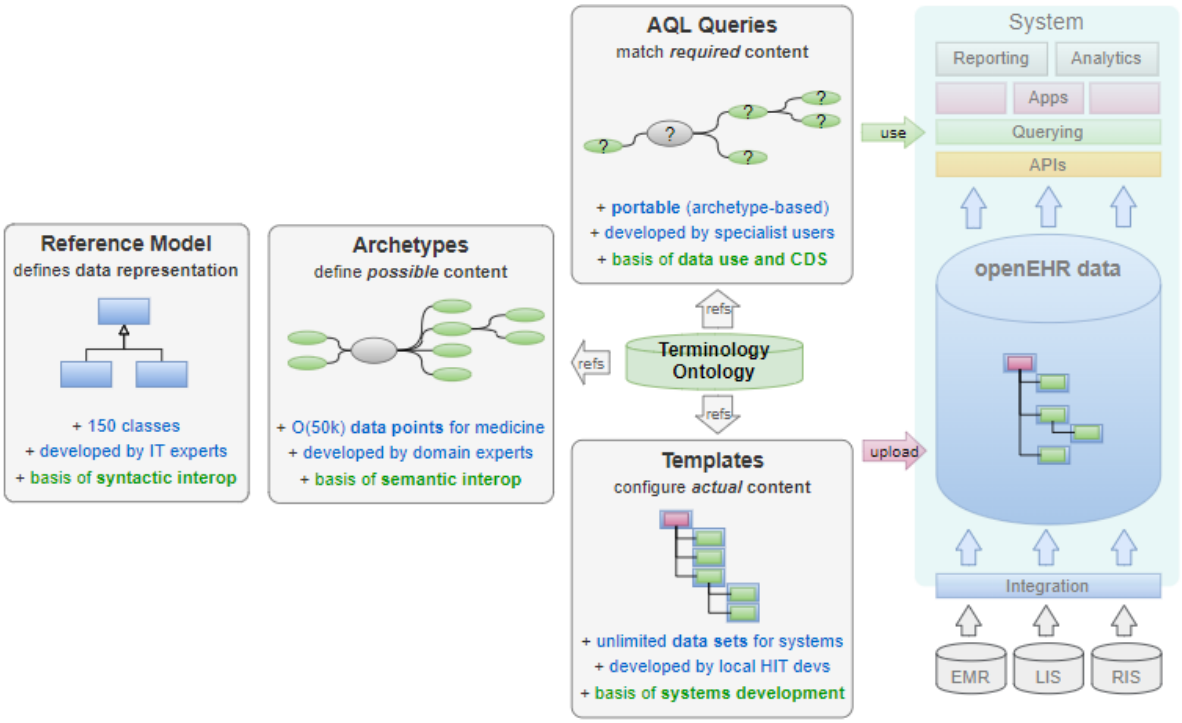

Figure 12. openEHR architecture [56]

The collection of information models, referred to as "reference models," is a key component of the openEHR standard [56]. They provide the immutable semantics of the EHR, EHR Extract, and the demographics model, while supporting data types, data structures, identifiers, and practical design patterns. The ENTRY classes' subtypes encompass OBSERVATION, EVALUATION, INSTRUCTION, ACTION, and ADMIN ENTRY, and are the most important classes defined by the standard. Another important class is the Instruction State Machine, a state machine that defines a typical model of medical treatments, such as prescribed medication, surgery, and other therapy. Archetype [57] is defined as reusable data point and data group definitions, i.e., content item that is reused in various contexts, or a template, which represents a data set specific to use-case. Template describes a use case-specific data set and consists of several archetype items' references. Semantic rules of archetypes cannot be violated. Templates are developed locally by developers in the form of GUI forms, message definitions and document definitions. Archetype Querying Language (AQL) [58] serves for writing queries for archetypes. openEHR archetypes are being used by the National e-Health Transition Authority of Australia, the UK NHS Health and Social Care Information Centre (HSCIC), the Norwegian Nasjonal IKT organization, and the Slovenian Ministry of Health. openEHR was selected as the base for the standardized EHR in Brazil [59].

openEHR was designed to provide a data platform with a primary focus on data durability and lesser on APIs and data exchange. openEHR employs approximately 300 more complicated archetypes than FHIR resources, which are intended to offer the greatest number of data components. Especially when contrasted to FHIR, which is geared for the easier, its volume necessarily adds a degree of complexity. Building patient-centered apps with an easy information exchange is made easier by the FHIR's API's simplicity and the availability of a simplified data model.

2.7. Electronic Health Record in Croatia

eKarton [60] is an Electronic Health Record system developed for Croatian national Health Information Service (HIS) which allows the healthcare professionals access to the patient's data and insight into their health record. All important patient information is available, including medical history, specialist opinion, all therapy, laboratory findings, etc.

The application itself was created using web technologies and data is available using Internet browsers and smart cards, not only by doctors in primary health care, but also by other users such as doctors in hospitals and emergency facilities and other authorized users who have a smart card issued by Croatian national health service, HZZO, and have an appropriate role that allows access to the system. Croatia's existing Electronic Health Record (EHR), i.e., eKarton in Central Health Information System (CEZIH) is depicted in Figure 13. The web application itself is shown in Figure 14.

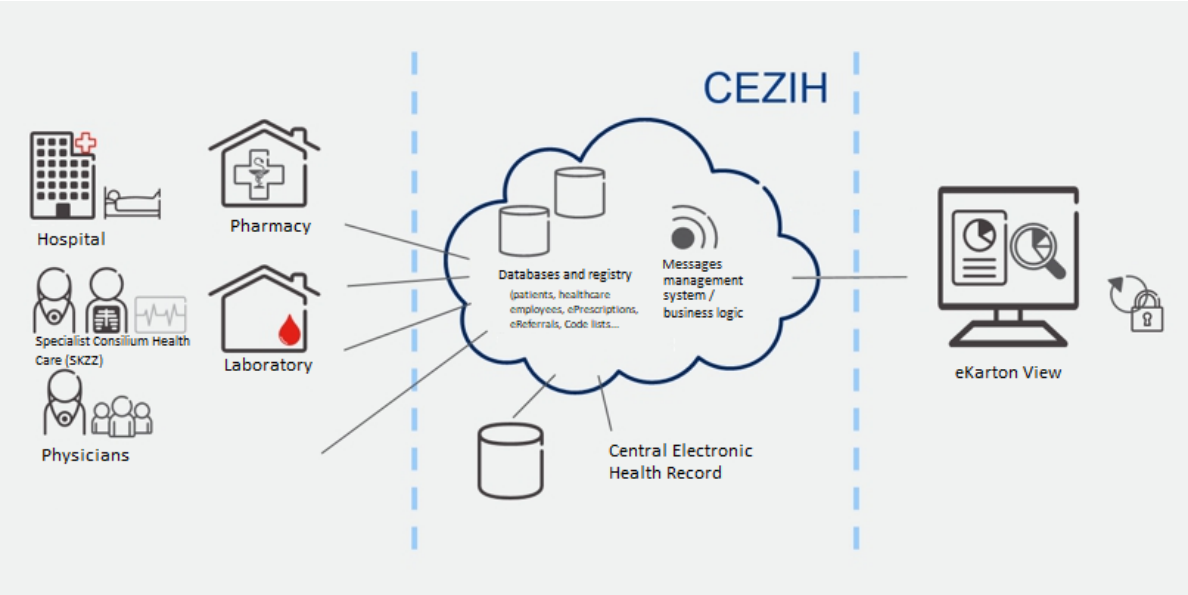

Figure 13. Electronic Health Record (eKarton) within Croatian HIS

Please note that the following figures do not provide real patient data, but instead are made up of dummy (test) data.

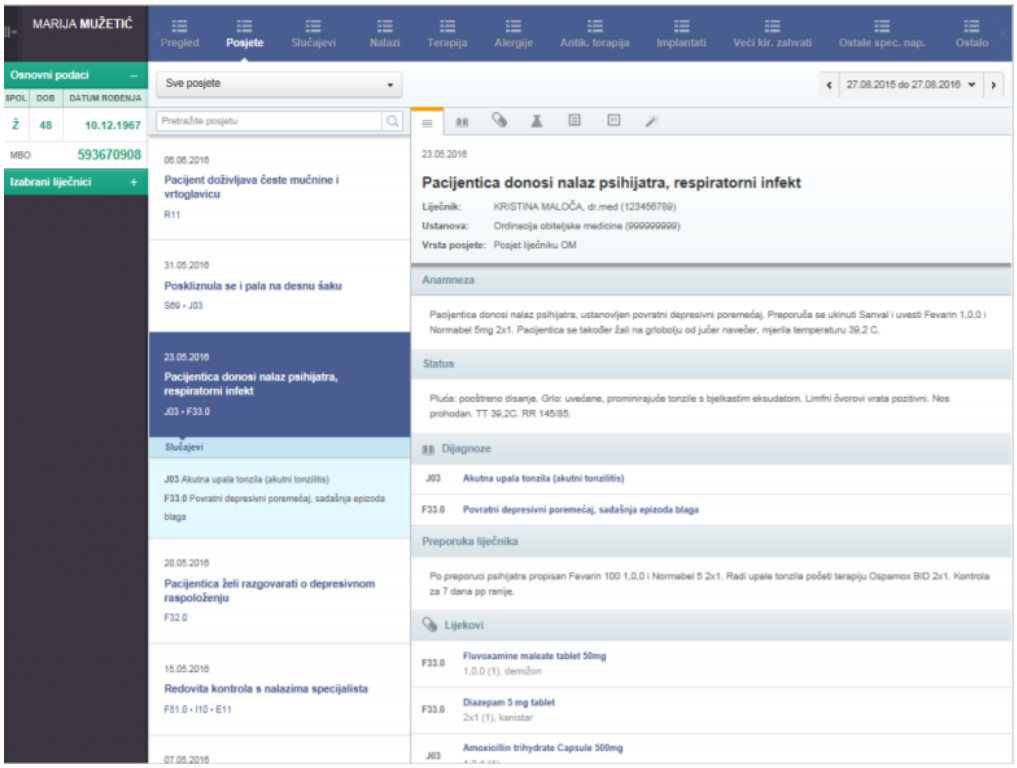

Figure 14. GUI for eKarton web-based application

Alongside the patient's general information, the physician can see in more detail past doctor visits, conditions, vaccinations, allergies, medications, surgeries, laboratory results, and additional notes selecting the appropriate icon (Figure 15).

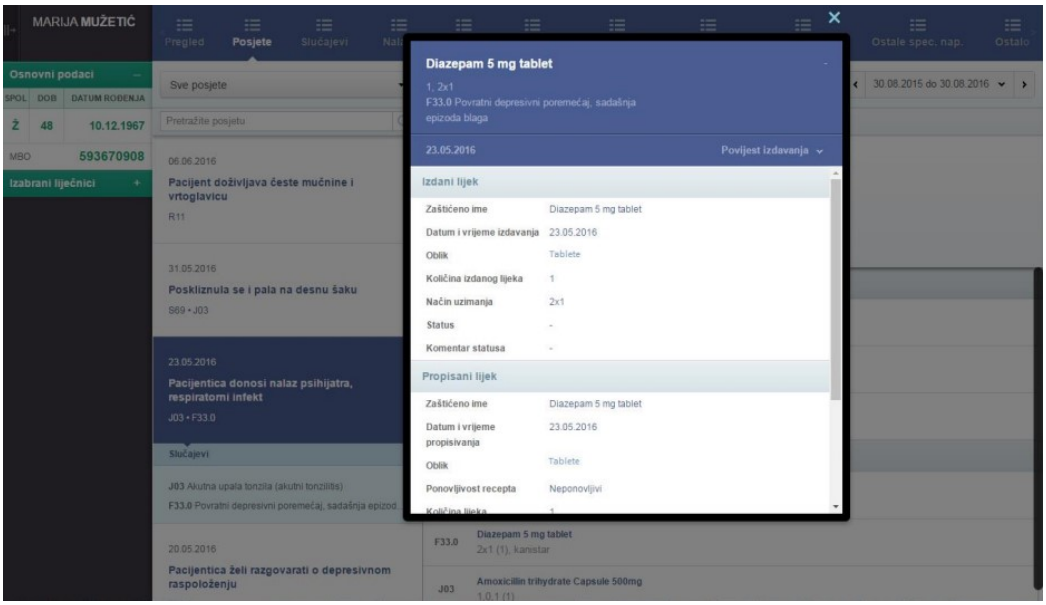

Figure 15. Detailed view of prescribed medication

Given the need for the patient's consent for a certain healthcare professional (whether a general practitioner, a chosen dentist, an ambulance worker, a pharmacist, etc.) to access eKarton data, these functionalities are available only to those healthcare professionals for whom patients have allowed access. Existing roles are chosen primary care doctor, chosen pediatrician, chosen gynecologist, chosen dentist, ambulance, specialists in hospitals and pharmacists. The level of access for each of the roles is determined by the patient. Others can only access basic patient identification data from selected physicians. No one can access other data and it is clearly stated that the patient has not given consent to access medical data.

Doctors utilize their Smart Card and PIN code to log into the system. Their role is then verified, and only if they have permission to see that specific patient's information will their data be provided. Individuals can grant specific doctor’s permissions, as shown in Figure 16.

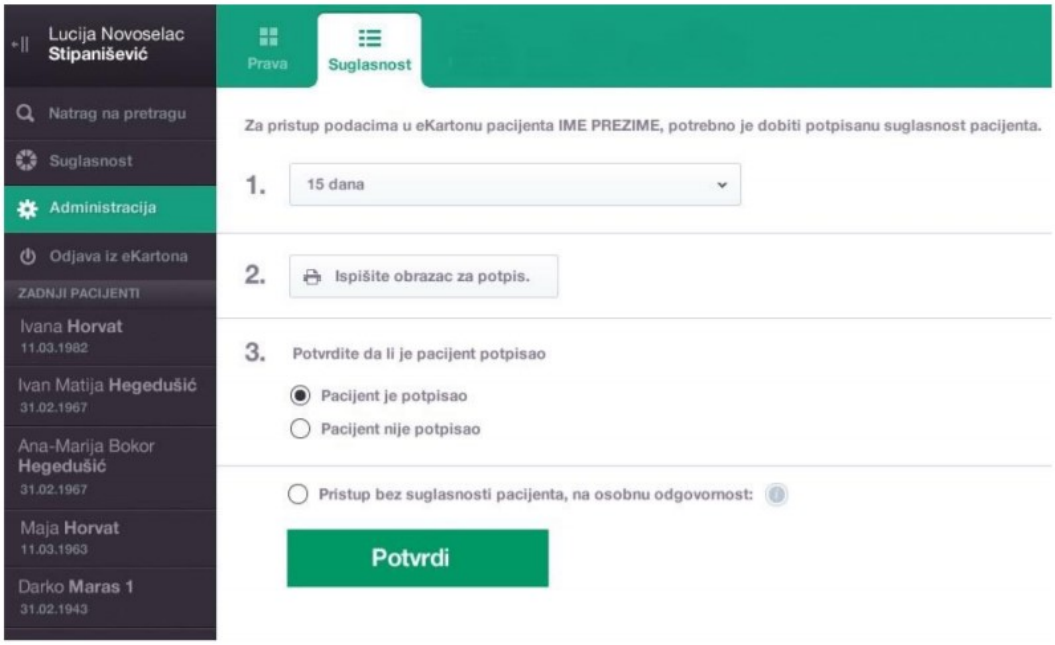

Figure 16. Patient has to authorize physicians' access to their EHR

The method that permits access in case of emergency exists (sometimes referred to as the "break the glass" procedure) in the event that the patient is in some way hindered from providing authorization and the physicians determine that access to the medical records is required. Using the Patient Portal (Portal Zdravlja) [61], the patient is able to see the history of users accessing their medical records, and change permissions, as pictured in Figure 17 below.

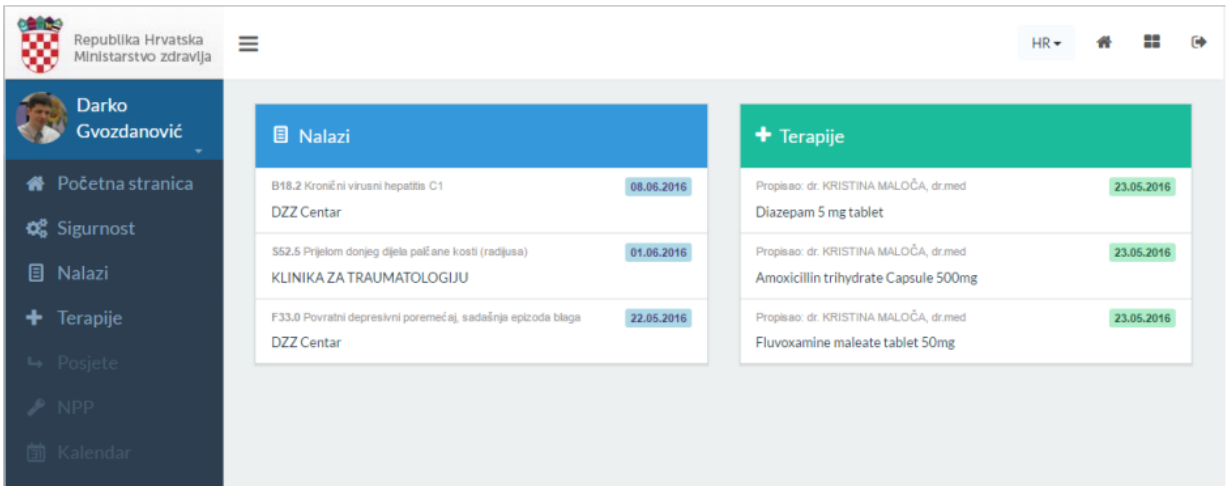

Figure 17. Patient Portal, PHR in Croatia's HIS

2.8. General EHR system architecture

Architecture of an EHR system is given in Figure 18. The implementation of an EHR system involves several key components, including databases, repositories, registries, audit logs, providers, and standardization (IHE, HL7).

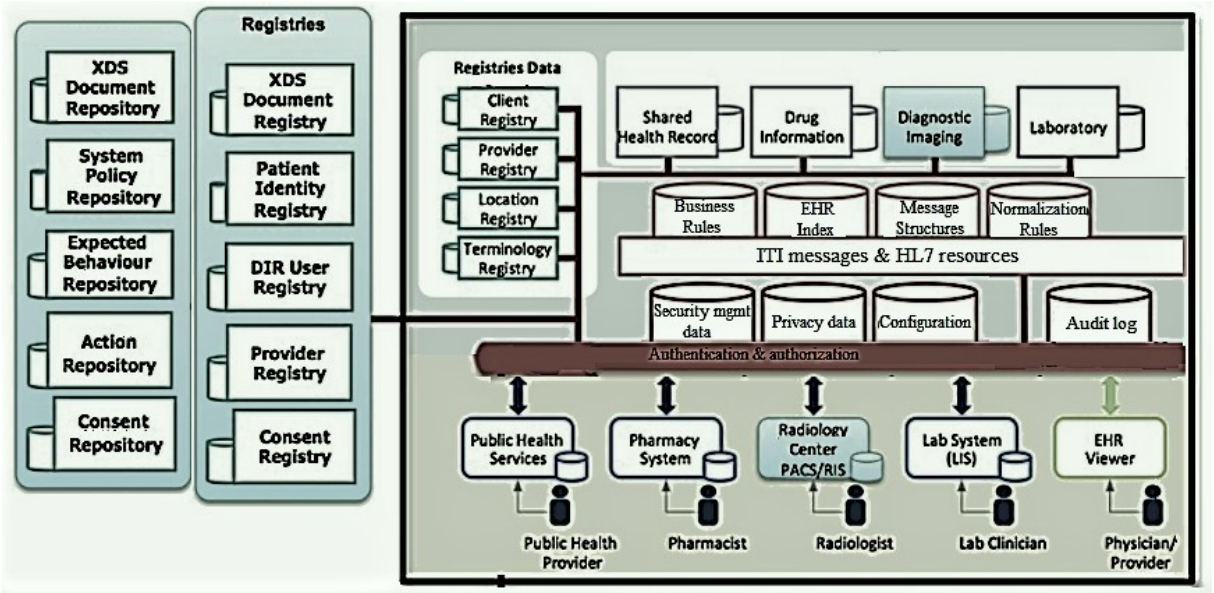

Figure 18. EHR architecture [38]

EHR systems typically involve the creation of one or more databases to store patient data. These databases may be centralized or distributed across multiple locations, and may use different technologies such as SQL or NoSQL databases. In addition to patient data, these databases may also contain information about healthcare providers, insurance plans, and other related entities. Data is stored in registries and repositories. In the context of EHR systems, a registry and a repository serve different purposes.

A registry is a database that contains information about a particular disease or condition, including data about the individuals who have been diagnosed with the condition. For example, a cancer registry would contain information about patients who have been diagnosed with cancer, including demographic data, details about their cancer diagnosis, treatment history, and outcomes. Registries can be used to monitor the incidence and prevalence of specific conditions, to evaluate the effectiveness of treatments, and to identify potential risk factors. A repository, on the other hand, is a centralized storage location for health-related data. In an EHR system, a repository might be used to store patient data such as medical histories, lab results, and diagnostic images. The repository can be accessed by authorized healthcare providers,

allowing them to view and update patient information as needed. In summary, a registry is a database focused on a particular disease or condition and contains information about patients with that condition.

A repository is a centralized storage location for health-related data, such as patient records, that can be accessed by authorized healthcare providers. In the context of electronic health record (EHR) systems, several types of repositories may be used to manage various aspects of the system's operation. Key repositories are:

- XSD Document Repository: An XSD (XML Schema Definition) document repository contains the schema definitions that define the structure of the data that can be stored in the EHR system. This repository ensures that data is stored in a consistent format, making it easier to search, retrieve, and analyze.
- System Policy Repository: The system policy repository contains policies and guidelines that govern the operation of the EHR system. This can include rules about data access, security, and privacy, as well as policies related to data retention and disposal.
- Expected Behavior Repository: The expected behavior repository defines the expected behavior of the EHR system under various circumstances. This can include scenarios related to data entry, data retrieval, system maintenance, and system failure. By defining expected behavior, the repository helps ensure that the system functions as intended and minimizes the risk of errors or inconsistencies.
- Action Repository: The action repository contains a log of all actions taken within the EHR system, such as data entry, updates, and access requests. This log can be used to track the activity of users within the system and to identify potential security or privacy breaches.
- Consent Repository: The consent repository contains records of patient consent for the collection, use, and sharing of their health data. This can include consent for treatment, research, and data sharing with other healthcare providers. The repository helps ensure that patients' privacy and autonomy are respected, and that their data is only used in ways that they have authorized.

Overall, these repositories help ensure that the EHR system operates in a consistent and secure manner, while also protecting patient privacy and autonomy.

In the context of EHR systems, several types of registries may be used to manage various aspects of the system's operation. An overview of these registries follows:

- XSD Document Registry: An XSD (XML Schema Definition) document registry contains the schema definitions that define the structure of the data that can be stored in the EHR system. This registry ensures that data is stored in a consistent format, making it easier to search, retrieve, and analyze.
- Patient Identity Registry: The patient identity registry contains information about patients who have been registered in the EHR system. This can include demographic data, such as name, date of birth, and contact information, as well as unique identifiers that are used to link patient data across different parts of the EHR system.
- DIR User Registry: The DIR (Data Interchange for Radiology) user registry contains information about users who have access to the EHR system's radiology data exchange capabilities. This registry helps ensure that only authorized users can access and exchange radiology data and can help track user activity to identify potential security breaches.
- Provider Registry: The provider registry contains information about healthcare providers who use the EHR system. This can include information about their professional credentials, areas of expertise, and contact information. The registry can help patients find and choose healthcare providers and can help ensure that providers have the necessary qualifications and training to use the EHR system effectively.
- Consent Registry: The consent registry contains records of patient consent for the collection, use, and sharing of their health data. This can include consent for treatment, research, and data sharing with other healthcare providers. The registry helps ensure that patients' privacy and autonomy are respected, and that their data is only used in ways that they have authorized.

Overall, these registries help ensure that the EHR system operates in a consistent and secure manner, while also protecting patient privacy and autonomy. By centralizing key information about patients, providers, and system policies, EHR registries can help improve the quality and safety of healthcare services.

The Business Logic layer of an EHR system is responsible for managing the business rules and processes that govern the operation of the system. The Business Logic layer defines different Business roles and permissions that users can have within the EHR system. This can include roles such as physicians, nurses, and administrative staff, each with their own set of permissions and access levels.

The EHR Index is a database that contains an index of all patient records in the EHR system, along with metadata about each record, such as the patient's name, date of birth, and medical record number. This index is used to quickly search and retrieve patient records when needed.
The Business Logic layer includes configuration files that define how the EHR system is configured, including settings for data storage, security, and privacy. These configuration files can be updated as needed to modify the behavior of the system. The Business Logic layer includes security management data that defines how the EHR system handles authentication, access control, and data encryption. This data helps ensure that the EHR system is secure and that patient data is protected from unauthorized access. The Business Logic layer includes privacy data that defines how the EHR system handles patient privacy, including policies around data access, sharing, and consent. This data helps ensure that patient privacy is respected and that the EHR system is compliant with applicable privacy regulations. It also defines the message structures used to communicate between different components of the EHR system. This includes the format of data messages, such as HL7 messages, and how they are structured and transmitted. The Business Logic layer includes normalization rules that help ensure that data is stored and retrieved consistently across different parts of the EHR system. This includes rules for data formatting, data validation, and data mapping.
Audit logs are records of all actions taken within the EHR system, including data access, modifications, and deletions. These logs are important for maintaining the security and privacy of patient data, and may be subject to regulatory requirements. EHR systems are used by a variety of healthcare providers, including physicians, nurses, and other care team members. These providers may access patient data from different locations, including hospitals, clinics, and other healthcare settings.
Overall, the Business Logic layer is critical to the effective operation of an EHR system. By defining business roles, managing patient data, and ensuring that the system is secure and compliant with privacy regulations, the Business Logic layer helps ensure that patients receive high-quality care while also protecting their privacy and confidentiality.
Standardization is important for ensuring that EHR systems are interoperable and can communicate with other systems. Two key standards used in EHR systems are Integrating the Healthcare Enterprise (IHE) and Health Level Seven (HL7). IHE defines a set of technical standards for integrating healthcare systems, while HL7 defines a set of messaging standards for exchanging healthcare information.
Finally, the implementation of an EHR system involves a complex set of components and stakeholders. EHR systems involve a range of stakeholders, including healthcare providers,

patients, healthcare organizations, government agencies, and technology vendors. Each of these stakeholders may have different goals and requirements for the EHR system, and it's important to consider their perspectives when designing and implementing the system. By considering each of these elements carefully, healthcare organizations can design and implement EHR systems that improve patient outcomes, streamline workflows, and meet regulatory requirements.

HL7 standards are the backbone of the National Health Information Service (HIS) in Croatia's eKarton, as well as in many other countries, including Finland, Norway, Sweden, Iceland [196], and India [197]. Health information exchange between databases and e.g. EHR in HIS is handled by the Messages management system (Figure 13). Messages, i.e. transactions, are defined in IHE Technical Frameworks.

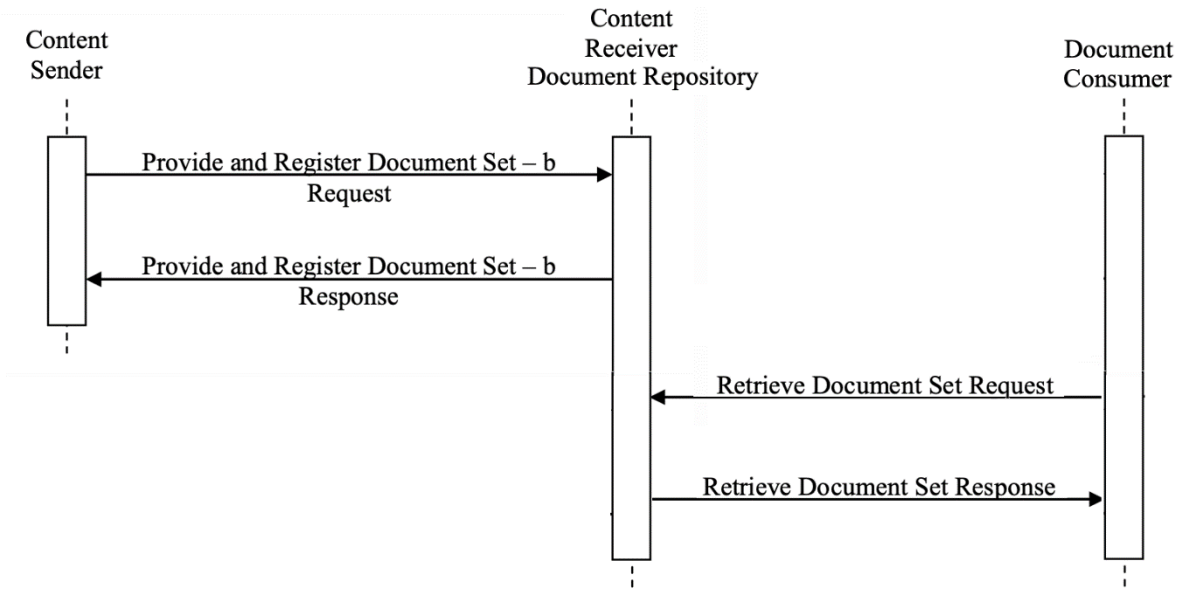

Figure 19. ITI-41 and ITI-42 messages [19]

IHE Technical Frameworks [44] outline how to put previously established standards into practice in order to achieve warranted medical information exchange, enable feasible and effective system integration, and provide the best possible patient care. The HL7 standard is the foundation for file sharing. A set of documents and related metadata are exchanged using the Provide and Register Document Set-b (ITI-41) transaction. Requests are unmarshalled, which means that the inbound data stream is converted into an HL7 Resource object (write operations) or message header parameters (search operation). When a message fails to unmarshall, an exception is thrown. Dependent on the actors and workflows employing the transaction, the documents and metadata may be handled, processed, and stored in order to be retrieved in the

future. Documents and related metadata are sent to a content receiver by a content sender. These events can be triggered by a human decision or as an automatic operation via application, wanting to submit a document to a Content Receiver, i.e., HIS repository. The transaction Retrieve Document Set (ITI-43) is used for retrieving a document from HIS repository. This flow is pictured in Figure 19. IHE Profile leverages HTTP, Web Services, IT presentation formats and HL7 Clinical Documentation Architecture (CDA). This enables it to handle HL7 Resources within the EHR system. The xds-iti43 component provides interfaces for actors of ITI-43 messages. The endpoint URI format of the ITI-43 component is defined as:

```
xds-iti-43://hostname:port/service[?params]
```

where hostname is domain name or IP address, service is path to the service, and params are optional parameters. An example of the exposed FHIR REST Service endpoint would thus be, e.g.:

```
http://ekarton.server.org:8888/IHE/xds/iti43
```

Current EHR implementations already use ITI messages in combination with HL7 standards, with existing defined and implemented service endpoints for various ITI messages. Ensuring the collected data exchanged with HIS is abiding to the messaging standards mentioned is covered in Chapter V.

3. Sensors and Data Collecting

Sensors are ubiquitous in today's world, and are used, among others, in healthcare, fitness, industry, and daily life [62]. A sensor is a device that "senses" something, for example level, temperature, flow, pressure, speed, or position. Sensors detect a change in physical, electrical, or chemical properties. Then, in response to the detected change, it produces an electrical output. Sensors vary from seismic, low sampling rate magnetic, thermal, visual, infrared, acoustic and radar, which are able to monitor a wide variety of ambient conditions like temperature, humidity, vehicular movement, light, pressure, soil makeup, noise level, object detection, mechanical stress level, dimensions, speed and many more. For example, they are used in smart homes for heat and light management, in cleaning robots or security system cameras. Proximity and motion sensors are used in vehicles as well as opening doors in shopping centers. In industrial processes, sensors are used to measure fluid levels in containers, to control temperature in manufacturing processes, or to monitor water quality and chemical changes [63]. In short, sensors help in managing day-to-day activities, ensuring safety, identifying, and reacting to emergency, monitoring environmental factors, enforce quality control of food and other products by monitoring moisture and temperature, help users track their fitness activities as well as help managing residential and business areas; from doorbells, automating lighting operations to household appliances for cleaning, boilers, and washing machines.

Sensors are the most commonly used monitoring technology. Sensors offer the collection of data from the environment within a short time period, and often connect to the cloud through the use of several communication and transport modes, e.g., mobile and satellite networks, Bluetooth, broad-based networks, low-energy wide-band networks, etc.

Historical chronology of the discovery and development of different sensors in context of materials (blue), sensor technology (orange), and biotechnology (green) is given in Figure 20 below.

In order to create entirely new classes of sensors, the creation of novel functional materials typically must be coupled with advancements in other domains. According to [64], materials are crucial in the development of sophisticated disposable sensing devices because they can reduce costs, have a positive influence on the environment, and improve functionality and usability.

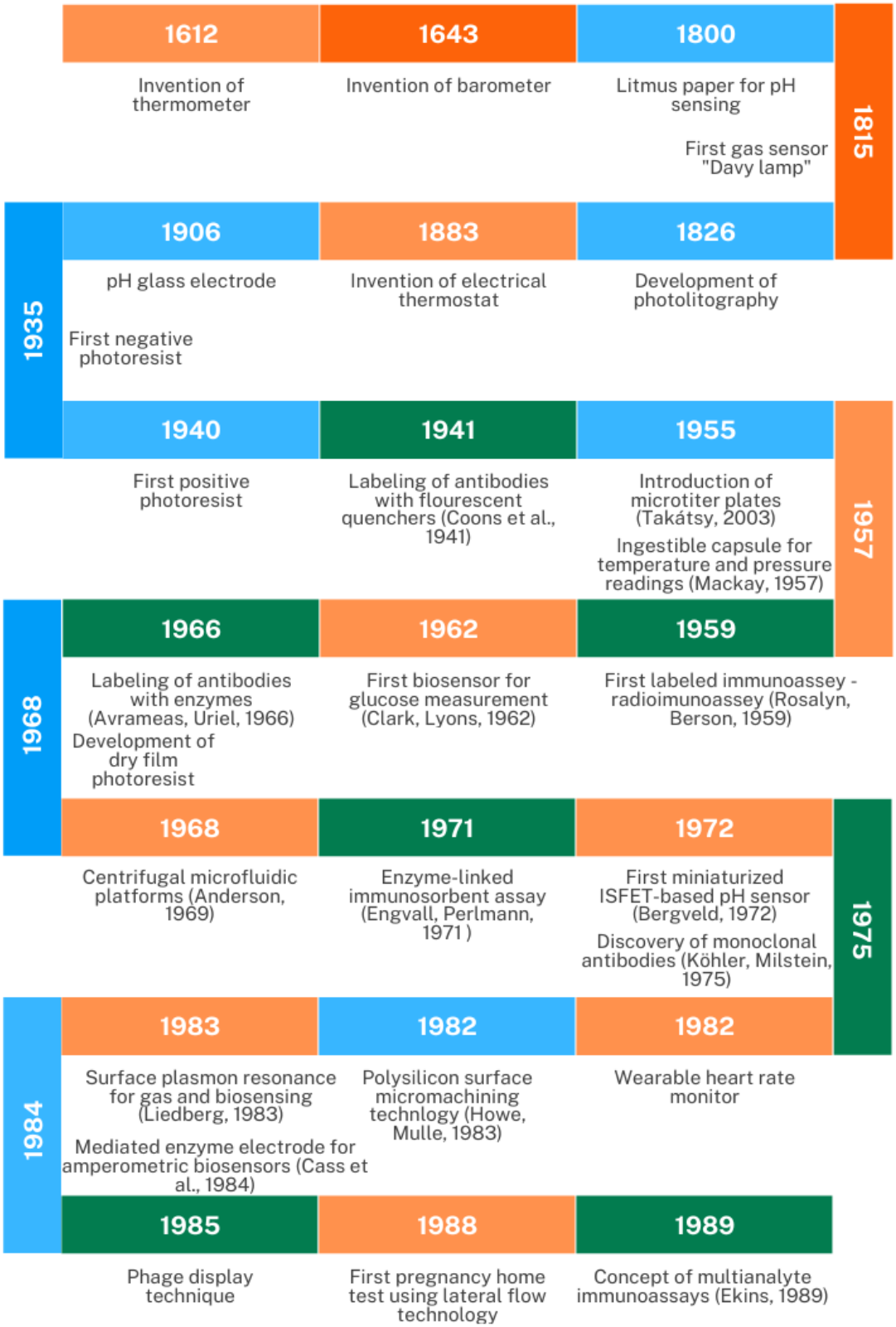

Figure 20a. Historical timeline of the development of different sensors in context of materials (blue), sensor technologies (orange) and biotechnology (green)

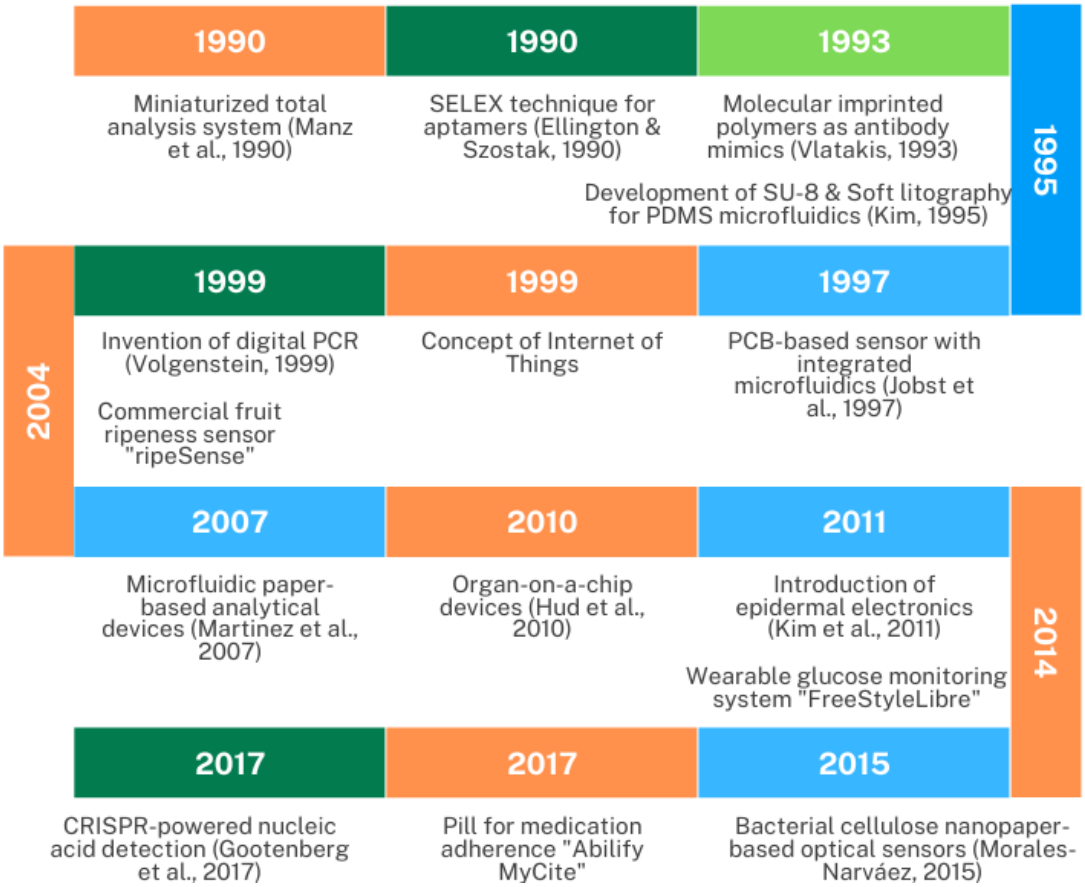

Figure 20b. Continuation of Historical timeline of the development of different sensors

Sensors can be classified into passive and active. Active sensors, such as thermocouple or a piezoelectric sensor, do not require an external power supply. Passive sensors require an external power source, such as an excitation circuit, to produce an electrical output. Examples of passive sensors include Resistance Temperature Detector (RTD) and strain gauge. Figure 21 shows active thermocouple and Figure 22 shows passive RTD.

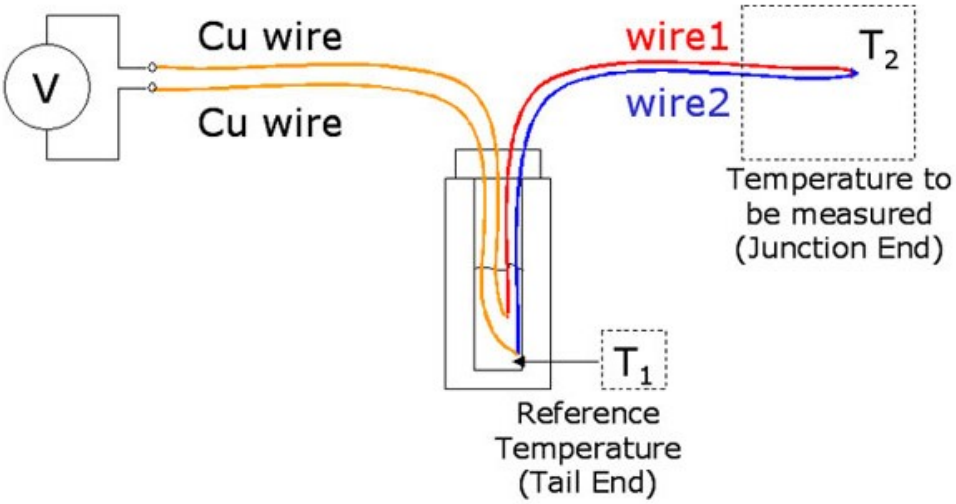

Figure 21. Active sensor – thermocouple

Thermocouple is widely used as temperature sensor in science, home and office thermostats, industrial processes involving kilns, gas turbine exhausts, or diesel engines, as well as flame sensor for gas-powered appliances or equipment. Thermocouple consists of two distinct electrical conductors that form an electrical junction. When increase in temperature occurs, thermocouple produces voltage, which can be interpreted to measure temperature. This is known as the Seebeck effect [66]. Commercially, these sensors are inexpensive and able to measure wide temperature ranges. However, they are known to be somewhat imprecise as system errors lesser than one degree Celsius are difficult to attain.

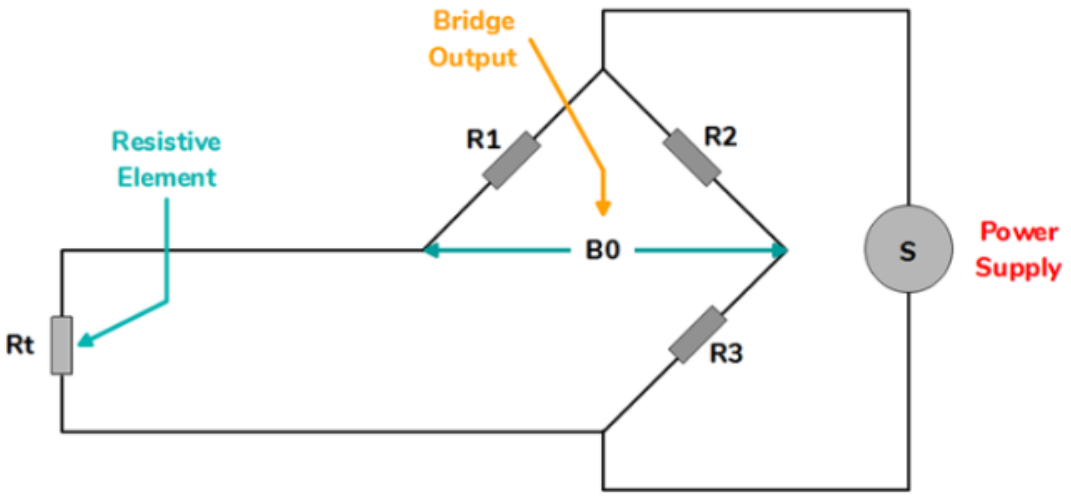

Figure 22. Passive sensor – 2-wire Resistance Temperature Detector (RTD)

RTD element usually consists of a thin wire wrapped around ceramic or glass core. Wire is usually pure nickel (Ni), copper (Cu) or platinum (Pt), with a known, accurate temperature-resistance correlation. Resistance of RTD changes with temperature. An external power supply or excitation circuit is needed to produce the change in voltage; electrical current is passed through the sensor in order to measure the resistance of the conductor, thus measuring the temperature. RTD elements are fragile, but they offer higher accuracy than thermocouples.

3.1. Sensor characteristics

Sensor characteristics regarding data quality and faults are:

- Sensor accuracy represents how close the output of the sensor is compared to the real value being measured. Accuracy is determined by testing and comparing the sensing system with a standard, known value or the readings of the sensors must be benchmarked against another system with already established very high accuracy.
- Trueness is a measure of systematic error and shows how close the average of the measured values is to the real value. Systematic error means that measured values of the same measurand will vary in a systematic, predictable way; they will diverge from the real value in the same direction or even amount.
- Precision is a measure of statistical variability or "random error" which is determined by the standard deviation, i.e., amount of variation or dispersion of a collected set of values. The lower the standard variation is, the values are closer to the mean or expected value. Consequently, the higher the standard variation is, the values are spread out over a wider range. Precision encompasses repeatability and reproducibility.
    - Repeatability is the level of agreement, i.e., how close the measures are among each other, when collected in independent measuring processes under identical conditions; meaning the same operator, material and instrumentation was used, and the measurements took place in a short period of time between them.
    - Reproducibility is the degree of agreement, i.e., how close the values are among each other, when collected in independent measuring processes under variable conditions, whether by using different instrumentation, material or operators, was conducted in vastly different time or both.

- Sensitivity is determined by the ratio of the difference between the output of the sensing system and the variation of the quantity being measured. Sensitivity can be linear (constant) or nonlinear (varied).
- Limit of blank (LOB) is the highest concentration of an analyte habitually detected when replications of a blank sample which contains no analyte are tested.
- Limit of detection (LOD) is the lowest analyte concentration which can be reliably measured against a blank sample (limit of blank) and at which detection is feasible.
- Limit of quantification (LOQ) is the smallest analyte concentration that can be detected with acceptable accuracy, meaning it has to meet predefined goals for bias and imprecision.
- Drift or operational stability shows how stable is the output signal of the sensing system with continuous and constant input in a long term. Drift can be caused by the changes in temperature and humidity or can happen due to the degradation of the sensor's components.
- Stability represents the sensor's capability to produce the same output signal over a period of time while measuring a standard, known value.
- Response time is defined as the time required for a sensing system to output a signal with a stable value.

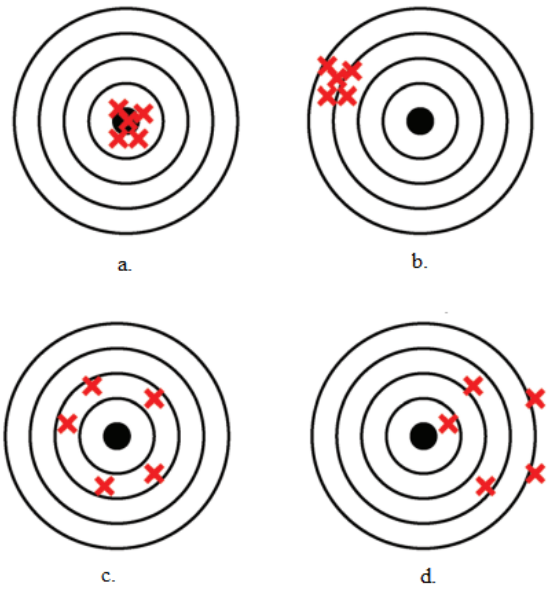

Figure 23. Examples of measured values with different data quality

Figure 23. shows in a. accurate and precise value; in b. precise, but not accurate value; in c. accurate, but not precise value; in d. not precise nor accurate value.

3.1.1. Classification of errors

By the use of measurements, it is impossible to ascertain a quantity's true value, which can be defined as the average value of an infinite measurement with average deviation approaching zero. Measured value is an approximation of true value, gained by the process of measuring. An error or fault can be defined as the disparity between the measured (approximated) and true value. The following categories of measurement errors exist, based on the sources from which the faults may arise (Figure 24).

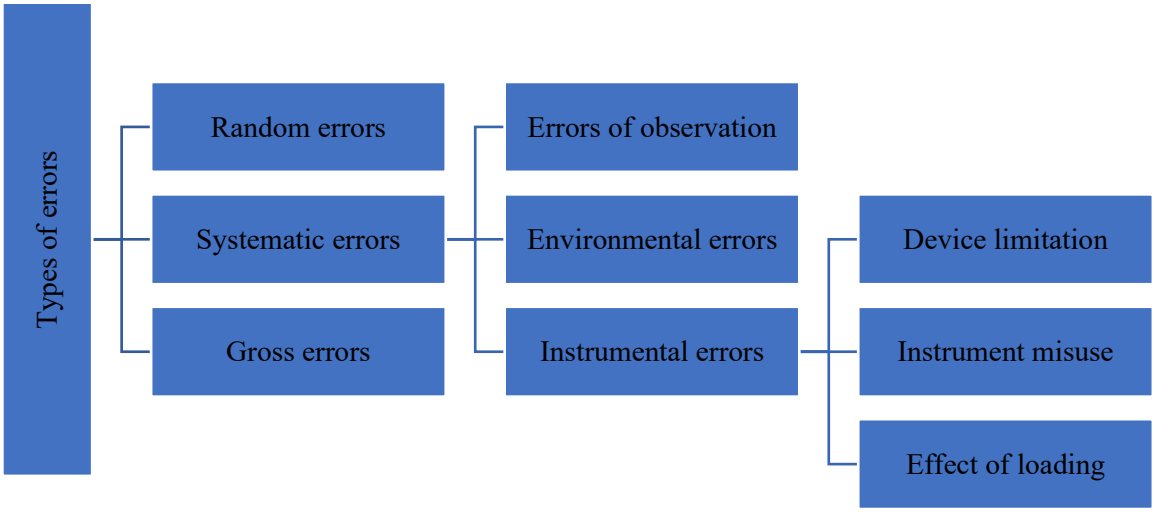

Figure 24. Types of errors

Systematic errors are errors not caused by chance but, instead, by repeatable system-inherent process. Error of observation is an error caused by imperfection in methods of observing a quantity, whether caused by instrumental or human factors. Environmental errors occur as a result of the measurement devices' environment. These kinds of faults usually occur because of temperature, moisture, dirt, vibration, or an electrostatic or magnetic field. Instrumental errors can occur due to the shortcomings of the instrument (e.g., friction in bearings of various moving parts, irregular spring tension), misuse or improper handling of the device, and “effect

of loading", i.e., after the instrument had been subjected to overload, the following measurements may be erroneous no matter the range.

Gross errors include errors made by humans when recording readings, documenting process variables, and computing outcomes. These readings happen as a result of the instruments' erroneous settings. There is no mathematical way to handle these faults. To minimize gross errors, when recording, calculating and taking readings, extreme caution should be exercised.

The cause for random errors in most cases remains unknown and, thus, they cannot be corrected by any method of correction or calibration. The frequency of measurement needs to be increased in order to prevent these faults. That implies that the same parameter should be measured frequently.

3.1.2. Sensor calibration

In order to obtain precise, accurate, and repeatable measurement results, sensors must be calibrated. Calibration is the widely used technique for eliminating or minimizing sensor bias. International Bureau of Weights and Measures (BIPM) defines the term calibration as an "operation that, under specified conditions, in a first step, establishes a relation between the quantity values with measurement uncertainties provided by measurement standards and corresponding indications with associated measurement uncertainties (of the calibrated instrument or secondary standard) and, in a second step, uses this information to establish a relation for obtaining a measurement result from an indication. [67]". The implementation of a test equipment monitoring procedure and the calibration of measuring tools are required by DIN EN ISO 9001 [68]. Due to use, aging, and environmental conditions, measuring instruments or equipment can lose accuracy, i.e:

- Incorrect zero reference – measuring instrument returns false reading when the actual value of a measured quantity is zero, e.g., the ammeter's needle returns non-zero value when no current flows. Since most modern sensors and transmitters are electronic devices, changes in external factors such as temperature, pressure, or time may cause the reference voltage or signal to drift over time.
- Error caused by Sensor Range Shift: Due to the aforementioned factors, the "sensor's range" may shift, or possibly the process's operational range has altered. For instance, a process may now function between 0 and 200 pounds per square inch (PSI), but modifications to its operation will need it to operate between 0 and 500 PSI.

- Error caused by mechanical wear or damage, in which case the instrument may require servicing or replacement.

As a result, sensors require periodic recalibration. The application determines the calibration schedule of the measuring instruments. Recalibration must be carried out periodically to maintain confidence in the traceability chain. The length of the intervals depends on a number of variables, such as the importance of the readings, the stability of the equipment, the frequency and usage pattern, or the threshold of required uncertainty. There can be several sensor calibration procedures employed depending on the measurement. Modern instruments can be calibrated manually or automatically. In calibration, a basic standard metric of accuracy is the Root Mean Squared Error (RMSE) [69].

In order for wearable sensors to be useful for measuring physical activity, they must be calibrated and validated against gold standard measures [200]. Calibration involves determining the relationship between the output of the wearable sensor and the actual physical activity being measured. This relationship can be affected by factors such as the type of physical activity, the position of the wearable on the body, and the individual wearing the monitor. Validation, on the other hand, involves comparing the output of the wearable monitor to a gold standard measure of physical activity. This can be done using methods such as direct observation or indirect calorimetry. Apart from highlighting the importance of calibration and validation for wearable trackers used in measuring physical activity, [200] reviews several studies that have attempted to calibrate and validate wearable monitors and conclude that the accuracy of these devices varies widely depending on the specific device and the activity being measured. They also note that there is a need for more standardized protocols for calibration and validation, in order to ensure that the results are comparable across studies. In another study [201], the authors tested three different wearable fitness trackers on a group of older adults with varied levels of ambulatory ability, using direct observation as the gold standard measure. Standardized protocols for calibration and validation, as well as careful consideration of the limitations of these devices, are important for ensuring that the results are accurate and meaningful. There are several methods for calibrating wearable monitors, including indirect and direct methods. Indirect methods involve using equations or statistical models to estimate energy expenditure or other parameters based on the wearable monitor's output. Direct methods involve measuring energy expenditure or other parameters using a reference method and then adjusting the wearable monitor's settings to match the reference method's measurements. Furthermore, [202] presents a study on the accuracy and reliability of wearable heart rate sensors. The authors aimed to develop a fully automated system that could calibrate and validate the performance of

such devices, with the goal of improving their accuracy and usefulness for clinical and research purposes. The data collected by the sensors were then compared to the reference measurements to assess their accuracy and reliability. A calibration curve was obtained by interpolating the experimental data, and a polynomial model was employed, resulting in a significant decrease in RMSE (0.81 to 0.33). The patent [203] describes a system for calibrating a set of sensors used to measure health-related parameters, such as body temperature and heart rate. The system includes a calibration chamber with a reference sensor and a set of test sensors, and a calibration circuit that applies a reference signal to the reference sensor and a test signal to the test sensors. The calibration process involves comparing the readings from the reference sensor and the test sensors and adjusting the test signals until the test sensors match the reference sensor, creating a baseline. The calibration circuit then stores the calibration coefficients for each test sensor, which can be used to adjust the sensor readings in subsequent measurements. Similarly, [204] presents a calibration method by using a second medical monitoring device to obtain conversion factor including first cross correlation that describes the correlations between measurements made by the two devices, such as a blood pressure monitor and a portable ECG device. In some cases, the second device can instead be a clinical "benchmark" device which a physician would use to calibrate the monitoring device. Using a data-driven approach improves the accuracy of the calibration.

3.2. Wireless Sensor Networks in healthcare

Wireless Sensor Network (WSN) comprises small electronic devices, sensors. Sensors can be used for monitoring areas, processes or objects, and tracking animals or persons. A sensor measures a physical phenomenon, such as temperature, humidity, light, pressure, or motion, among others. Output of a sensor is quantifiable data that can then be further processed and interpreted by a person or a machine. Wireless sensors have a number of limitations as a result of their low cost and low complexity, including a small transmission range, limited computing and processing power, lower reliability and data transfer speeds, and a finite amount of energy. Thus, WSNs must be designed with the goal to circumvent these limitations, for example, by exploiting the synergy among multiple nodes.

The type of sensor deployed, the area and extent of deployment, the computational needs, and the QoS requirements can vary greatly, depending on the application. Common use cases include agronomic industries, tracking resources in heavy industries and research laboratories,

manufacturing processes, monitoring of environmental conditions for forecasting and observation, monitoring remote and inaccessible areas, emergency and disaster-relief systems, automated processes, smart homes, as well as fitness and health-related applications.

Sensors have a variety of uses in the field of eHealth, including telemonitoring of a person's physiological data, general monitoring of patients, diagnosing the patients, and the administration of medications in medical facilities [70]. The use of sensors in both fitness and healthcare can be very beneficial in identifying internal complexity and disease causes, which would ordinarily be quite challenging. Some of the examples of sensor applications in healthcare include the Smart Sensors and Integrated Microsystems (SSIM) used for monitoring the levels of glucose, sensing allergens and monitoring respiration rate, tracking of blood pressure and monitoring of cardiovascular diseases, as well as general monitoring of physical and mental health.

Initial applications consisted of a single wireless sensor that would serve for monitoring and predicting of activities or health conditions. Some examples include walk analysis using sensors embedded in wearables [71] or flooring within a smart home [72], as well as and daily activity tracking from wearable motion sensors or sensors placed in the surroundings [73]. While applications based on video and/or audio inputs can produce gigabytes of data, the volume of data that may be stored is restricted to megabytes [74]. The majority of the time, processing was done centrally and in batches that were retrospectively processed offline.

Technological advancements regarding sensors facilitated continuous monitoring using a number of sensors, each of which is accountable for generating inference, whether from wearables or ambient sensors. Thus, the concept of an agent was introduced. Agent is a processing entity capable of both detecting and acting in order to accomplish an aim. These activities may depend on an agent's autonomy in interacting with the environment or on the collaboration with other agents. Therefore, a higher-level of reasoning was needed than in the initial devices to integrate the outputs from a variety of sources. By combining multisensory data and/or offering a concept of context-awareness, the goal was to minimize the uncertainty of predictions. One example of this is monitoring system or sleeping disorder [75], employing wearables, ambiance sensors, and video recordings to determine the most significant sleep-related events. Furthermore, [76] describes a fall detection system being used to recognize environmental risks by understanding the context of a fall using data from wearable motion and ambient vision sensors as well as electricity consumption (devices and lights switched on and off).

Novel approaches try to integrate various sources of clinical knowledge with continuous tracking of a person's health. Current research aims to incorporate intelligent agents that leverage modern technology (stream processing, data mining, genetical and multiomics data coupled with the pervasive sensing abilities of the previous applications. Thus, the agents are capable of retrieving data from a range of domains, such as medical research, EHR, and data collected in laboratories (such as genomes, proteomics, and metabolomics). These systems assess a patient at a system level, taking into consideration all contributing aspects, through the efficient merging of multimodal data [77]. This will aid in making decisions based on the most recent biological and health informatics findings. The ability to combine knowledge from several sources has the potential to greatly advance and personalize diagnosis and treatment. One of the examples is a non-invasive 3D printed colorimetric early wound-infection indicator [78] that can provide an early warning of infection before it has progressed into a chronic state by responding to a rapid, aerobic microbial colonization by changing the color. (Figure 25). Another example is a degradable silk-based subcutaneous sensor measuring blood oxygen levels [79]; the small circular silk film oxygen sensor placed under the skin glows purple when exposed to UV light and oxygen. A detector can then determine the level of oxygen by the brightness and duration of the purple glow (Figure 26).

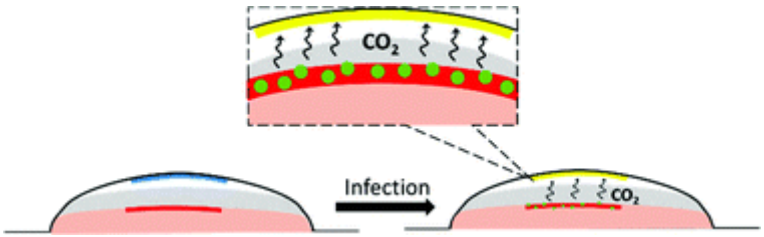

Figure 25. Non-invasive 3D printed colourimetric early wound-infection indicator [78]

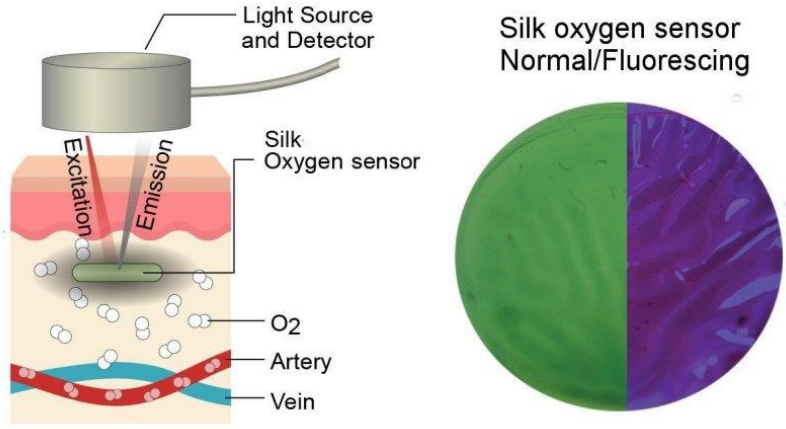

Figure 26. Degradable silk-based subcutaneous oxygen sensor [79]

A wearable sensor with a gold mesh that can measure different biomarkers or substances to perform on-body chemical analysis using Surface-Enhanced Raman Scattering (SERS) technique was developed [80]. SERS is a method of detecting the presence of a chemical indirectly by using laser light and a specialized sensor (Figure 27).

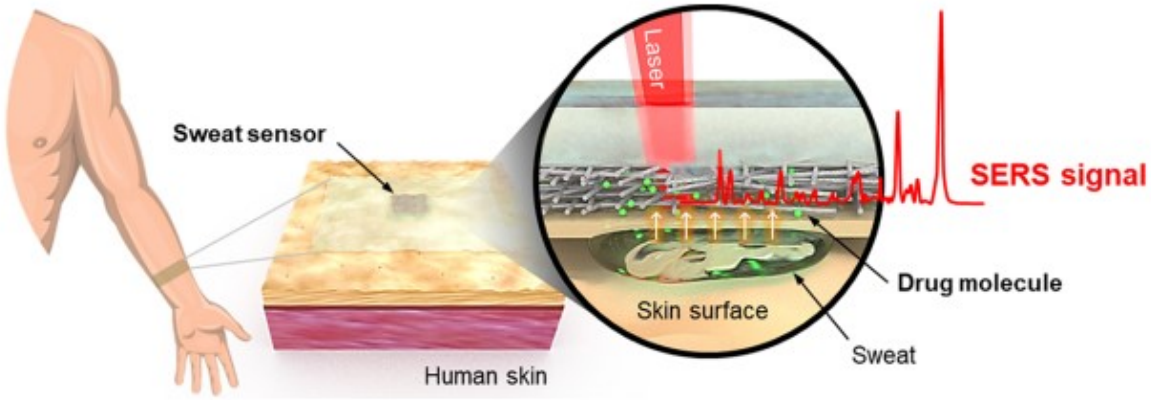

Figure 27. SERS detection of drug molecules in sweat [80]

Finally, review [81] classifies sensors in the healthcare domain into physical, electrochemical, optical sensors, and magnetic sensors. It also presents a full taxonomy of sensors deployed (Figure 28).

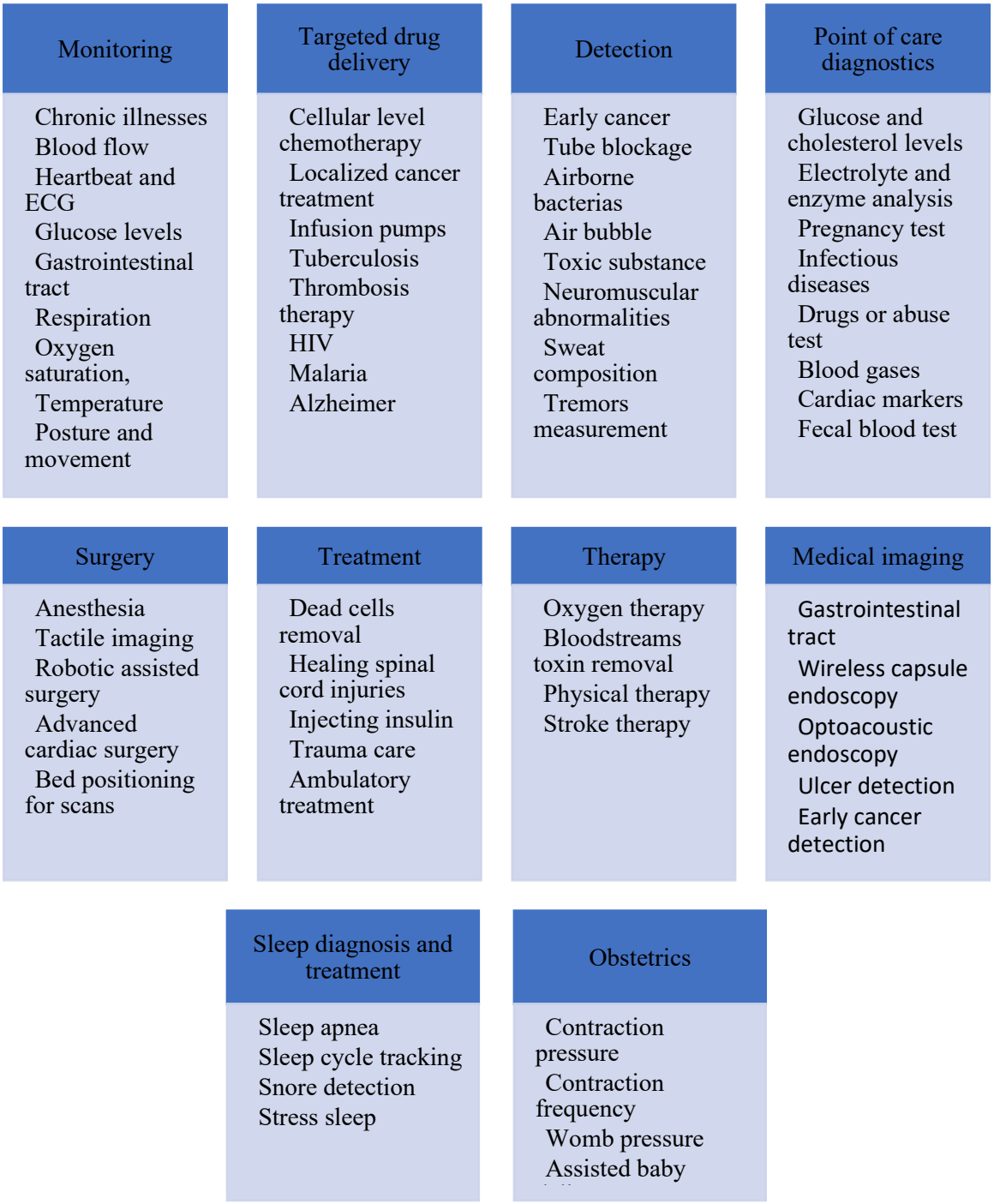

Figure 28. Taxonomy of sensors in healthcare

Personalized health data, such as vital signs, collected continuously and unobtrusively could potentially provide the physicians significant assistance in monitoring or diagnosing the patient. In this research, two datasets have been used:

- PMData dataset [146] made available via Simula Open Datasets,
- OxyBeat [147] dataset gathered for the purpose of this study to ensure a more robust use-case scenario by adding a number of additional data types, with an

emphasis on COVID-relevant data types (body temperature and oxygen saturation).

Both datasets were collected using Fitbit Versa activity trackers, i.e., devices equipped with sensors used to track fitness and health related metrics. Using the same wearable device, it is expected that the data will be largely comparable. One factor that can affect the quality of sensor data is the accuracy of the device. Some sensors may be more accurate than others, depending on their design, construction, and calibration. For example, a sensor that is designed to measure heart rate may have a higher level of accuracy if it uses advanced algorithms to filter out noise and interference from other sources. Similarly, a sensor that is designed to measure temperature may have a higher level of accuracy if it is calibrated regularly to ensure that it is measuring the correct temperature range. Using the same manufacturer and model of the device ensures that the sensors collecting the data are equivalent and the data is subjected to the same proprietary processing. Another factor that can affect the quality of sensor data is the way the device is worn. Some sensors need to be worn close to the skin, while others can be worn on clothing or other accessories. The way a sensor is worn can affect its accuracy, as it may be subject to interference from clothing, movement, or other factors. For example, a heart rate sensor that is worn too loosely may not be able to detect the wearer's heart rate accurately, while a sensor that is worn too tightly may cause discomfort or even injury. User behavior is also a factor that can affect the quality of sensor data. Users may behave differently when wearing a sensor, which can affect the accuracy of the data collected. For example, a user who is nervous or anxious may have a higher heart rate than usual, which can affect the accuracy of a heart rate sensor. Similarly, a user who is very active may produce more movement than usual, which can affect the accuracy of a motion sensor. To ensure that sensor data is accurate and reliable, it is important to choose high-quality sensors that are calibrated regularly, worn correctly, and used in a consistent manner. It is also important to take user behavior into account when analyzing sensor data, and to use advanced algorithms and statistical techniques to filter out noise and interference from other sources. In conclusion, while there are many factors that can affect the quality of the data, the two data sets being used were collected by the same wearable device model and thus are largely comparable.

3.2. Wearable Activity Trackers

Wearable activity trackers, commonly known also as fitness trackers, are devices worn by users, usually in the form of a wristband or a smartwatch which are used to collect health-related data. Research on the global wearable technology market [82] reports the wearable technology market was valued at 115 billion US dollars in 2021 and is projected to reach 380 billion US dollars by the end of 2028. Wrist-worn wearables account for 30% of the global market.
Something similar to modern fitness trackers first appeared in 1965 when Dr. Yoshiro Hatano, professor at Kyushu University of Health and Welfare, Japan was researching ways to combat obesity. Dr. Hotano invented Manpo-kei (Figure 29), a 10,000 steps meter. His later published research [83] and [84] states that 10,000 steps is necessary daily activity which maintains the body healthy, supposing the caloric intake is balanced properly. Today's fitness trackers still use 10,000 steps as a benchmark goal.

Figure 29. First modern pedometer, manpo-kei [84]

In 1980, Finish company Polar patented wireless heartbeat rate measurement, invented previously by the company's founder, Seppo Säynäjäkangas. The world's first wearable heart rate monitor, Sport Tester PE 2000 (Figure 30) was launched in 1982 by the same company.

Figure 30. First wearable heart rate monitor [92]

In 2006, mobile phones started incorporating 3D accelerometers which measure movement and vibration on three different axes: up and down, side-to-side, and front to back. Nokia's 5500 Sports was the first mobile phone that implemented this and, thus, was able to track the user's physical activity more accurately.

Today, modern wearable fitness trackers contain various sensors and keep track of various health-related parameters. Table 3 lists sensors which can be commonly found in modern wearable activity trackers.

Table 3. Most common sensors in wearable activity trackers

3-axis accelerometer	Senses motion and movement on three axes, using measurements of velocity and position. It senses inclination, tilt and orientation of the body. Ubiquitous in fitness trackers.
Gyroscope	Measures orientation and rotation, used for navigation and measuring angular velocity. A 3-axis gyroscope combined with 3-axis accelerometer provides a "6 degree of freedom" motion tracking system that's used in the majority of fitness trackers as it is useful when tracing workout motions.
Temperature sensor	Senses temperature. Combined with motion readings, it measures physical activity.
Bioimpedance sensor	Measures galvanic skin response (resistance to small electric current). Used to interpret activity and collect heart rate data.
Optical sensor	Preferred way to measure heart rate using light.
Altimeter	Measures altitude by using pressure sensing.

The data collected via sensors is then processed using manufacturer's proprietary algorithms in order to generate more detailed information, e.g., the data collected by the 3-axis accelerometer is used to calculate how many steps the user has taken, what was their speed, and at what pace, as well as the calculation of how many calories were likely burned. Most common measured metrics are given in Table 4.

Table 4. Most common metrics measured or calculated by wearable activity trackers

Steps taken	3-axis accelerometer
Distance covered	3-axis accelerometer and gyroscope
Floors climbed	Altimeter. This metric is also used for calculating calories expenditure and workout.
Heart rate	Optical sensor uses light and reflection to check the speed of blood flow on the wrist.
Body temperature	Measures temperature, also used to calculate physical activity and menstrual cycle as well as detect health issues (e.g., fever).
Oxygen saturation (SpO2)	Deoxygenated blood in veins is of a darker red color than the oxygen-filled blood in the arteries. Sensor measures relative reflection of red and infrared light. SpO2 value is estimated taking into account heartbeat rate as well.
Exercise time and calories burned	Both are calculated taking into account steps taken, distance covered, movement, velocity, and altitude, as well as heart rate, and body temperature.
Sleep duration and sleep quality	Estimated by monitoring body movements, changes in heartbeat rate, body temperature, and oxygen saturation.

3.3. Activity trackers and Quality of Data (QOD)

Medical devices and related services need to consistently fulfill the regulatory requirements specified in standard ISO 13485:2016 [85]. This encompasses design and development, manufacture, storage, as well as distribution, installation, and maintenance of the device. Standards regarding diagnostic equipment, including medical monitoring equipment, medical thermometers and related materials are under ICS code 11.040.55, such as ISO 80601-2-56:2017 [86] for clinical thermometers for measurement of body temperature. Testing, sampling, and calibration are generally covered by the standard ISO 17025:2017 [87]. Proper calibration results in reliable measurement traceability and accurate compliance to the standards. Nevertheless, regular usage of medical measurement tools is prevalent. Because the quality of medical apparatus directly affects patient's life and health, thus medical devices should be examined with more scrutiny. More focus should be given to calibrating medical measuring devices because they are an integral part of treatment. This is particularly true for devices that come into direct touch with patients, for equipment whose proper usage is essential to the patient's health, or for equipment that serves as a diagnostic tool for additional therapies. Due to the complexity of most medical measuring equipment, multiple distinct physical characteristics must be controlled during calibrations (e.g., voltage, resistance, time, temperature, pressure, mass, etc.). Regular calibration is required to check for accuracy and precision. If any faults are discovered, corrective action must be done, and any negative impacts must be assessed and recorded.

The data must be accurate, precise, and error-free in order to be used in a formal medical practice. In this context, it is necessary to revise how to ensure the quality of data collected by fitness trackers.

Wearable sensors, if worn incorrectly, e.g., too loosely, can report inaccurate readings due to “contact” problems between sensors and the skin. Fitness tracker should not be worn on the wristbone, as it prevents it from laying flat against the skin. This can cause additional light to enter and modify any readings made via optical sensor (e.g., heartbeat rate). It should not be worn too tightly, nor too loosely; ideally, two fingers should be able to fit between the wrist and the band. In order to ensure the necessary level of accuracy, wearables must be worn appropriately. Furthermore, all sensors have slightly different threshold levels which makes it improbable that two fitness trackers combined would result in exactly the same measurements of health-related data for the same individual, i.e., in a situation where a person wears multiple different activity trackers on the same wrist, it is likely that the results yielded by the different

trackers will vary in measurements. That, however, does not indicate they are at fault; but rather indicates that the sensor's readings differ slightly or that different manufacturers use different proprietary algorithms to calculate health-related metrics. The study [88] featured twenty individuals utilizing five distinct wearables. Nevertheless, it found that the reported distance and steps taken by several devices worn concurrently by the same participant can differ up to 26% depending on the device worn. On various devices, there was low correlation between the number of calories expended and the number of steps taken. This demonstrates how the computations depend greatly on the wearable itself and the manufacturer's proprietary algorithms. Because of this, it is challenging to use such metrics as a precise determinant of health issues, making it impossible to use them in a professional medical environment without some type of data cleaning procedure.

In paper [89], to increase the accuracy of both step and energy expenditure (EE) estimation, comparison and evaluation of regression-based models was performed. Seven fitness wearables were utilized to collect data, while two other devices served as references. In three separate iterations, twenty young adults wore all of the devices at once. Each of the devices' EE measurements and five of the devices' step measurements improved as a consequence of the creation of regression models for all the devices using reference data. Similarly, [90] examines several data-driven methods for cleaning eHealth sensor data from IoMT devices and offers recommendations for improving them to increase the acquired data's accuracy in preparation for inclusion in a formal EHR.

Thus, it is crucial to take into account that, while fitness trackers worn by an individual can serve as a reference point for the physical activities of an individual, careful interpretation is needed. It is typical for the numbers for steps taken, heart rate, and other metrics to vary amongst trackers. The statistics will make sense over time as long as the person uses the same activity tracker, so long as there is a certain continuity. However, this means special attention should be given when choosing the wearables considered for use in medical context, taking into account the lack of rigorous calibration that medical equipment is subjected to, as well as potentially decreased accuracy due to aging of the sensors. Numerous studies on commercially available wearable heart rate monitors were done. Study [91] measured the accuracy of wrist-worn commercial heart rate monitors compared to ECG on 25 persons, resulting in correlation coefficient with ECG reaching up to 0.92 (0.903-0.934 depending on the participant), which is very promising. Similarly, [92] reports a coefficient of variation (CV) (%) of less than 5% across different fitness activities for some models. Another study [93] assesses accuracy of fitness trackers pertaining to heart rate and energy expenditure, using Bland-Altman analysis,

correlational analysis, and error bias, resulting in mean relative error (RE, %) of -3.3% to -4.7%. Common conclusion is that tested wearables have shown acceptable heartbeat rate accuracy with a small negative bias. Out of the brands currently available on the market, the five most utilized in research projects are Fitbit, Garmin, Misfit, Apple, and Polar. Fitbit has been used in twice as many validation studies as any other brands and is registered in Clinical Trials [94] studies ten times more frequently than other brands. Fitabase library [95], shows numerous clinical trials and systematic reviews of used Fitbit wearables which resulted in 992 publications at the time of writing. Additionally, systematic review for validity and reliability of fitness trackers [96] reports higher consistency and correlation coefficients (CC) on Fitbit devices in various studies for steps taken, energy expenditure, sleep time and sleep efficiency. Wearable sensors possess impressive accuracy, but their performance can be enhanced by implementing calibration procedures periodically and employing data cleaning techniques to remove instances of low confidence data that may compromise the overall accuracy of the sensor readings, thereby facilitating more precise data capture.

4. System model for data cleaning and transformation

A step closer to ensuring continuous healthcare access would be achieved if EHRs were equipped to store and interpret data collected by sensors. This would result in standard-compliant personalized medical services. However, there are a number of faults and errors that might affect sensor data, which can further result in imprecise or even erroneous and deceptive answers. For the data to be used in a formal EHR, it is crucial to ensure the quality of the data gathered from sources such as wearables. Wireless sensors may be employed for remote monitoring, object tracking, and a variety of other applications in a broad range of contexts. WSNs can be utilized in eHealth for telemonitoring of personal physiological data as well as monitoring patients, diagnostics, administering medication in hospitals, and tracking of the internal processes of extremely sensitive bio-fluids. The utilization of sensors in healthcare and medicine may be incredibly beneficial in identifying internal complexity and disease causes, which might otherwise be quite challenging given that these phenomena are linked to the patient's interior anatomy. Notable instances of wireless sensor devices in healthcare application include monitoring blood sugar levels [97], cancer detection [98], monitoring organs [99], as well as overall health supervision. The primary purpose of WSNs is data collection, that is accomplished via periodic, distributed, and cooperative sensing and transmission actions by sensor node [100]. The network-collected sensor data expands exponentially and exhibits big data traits. Data integrity protection technologies have not caught up with the growth of WSN, despite the prevalence of these networks increasing [101-103]. One major contributing factor to this is an absence of in-depth understanding of the various fault types that might arise in WSNs. In general, there are various potential sources of faults in sensor data (i.e., data points do not accurately represent the physical phenomena measured by the sensor) [104] [105]. Before designing an efficient process for fault detection in WSN, a suitable model of the faults in the system is necessary. The following potential faults are defined by the model of data faults as stated in [106]: Discontinuous or intermittent faults whose occurrence is sporadic and discreet. If the erroneous readings occur regularly (at a frequency rate above the threshold), the sensor is malfunctioning. If not, faults are viewed as random. Continuous or regular if a sensor consistently gives off erroneous data over the course of the observation period and patterns in the form of a function can be seen. These flaws may be further broken down into:

- bias (the error function is a constant with either a positive or negative offset)

- drift (the deviation of data can be described by a learnable function, e.g., polynomial change).

Such faults may appear as noise (random undesired variation in data that generates ambiguity in data), erroneous measurements, or missing values and may be brought on by a variety of external factors including environmental impacts and perturbations or by flaws in the equipment. Because of its low cost and sensors' limited resources, many hardware-related flaws emerge (WSN devices are not able to store a large amount of data due to memory constraints, processor constraints make operations slow and short battery life is due to battery capacity limitations) [105] [106], and variation of Link Quality Indicator (LQI), that acts as a gauge of a communications channel's suitability for accurate signal transmission and reception [107]. This could also result in answers that are imprecise, inaccurate, or misleading. The most recent data quality study is reported in [108]. Furthermore, [109] outlines 24 attributes or properties as criteria of health-related data quality from different sources emerging in the US, Canada, the UK, Australia and WHO. The ten criteria listed below are indications of data quality according to the American Health Information Management Association (AHIMA) data quality model:

- accuracy – data shouldn't have errors, noise or missing values
- accessibility – data needs to be available for access,
- comprehensiveness – clearly documented what the collected data represents,
- consistency – data is reliable and without contradictions,
- currency – the data is up to date, outdated data is clearly marked or removed,
- definition – all data elements must be singularly defined,
- granularity – the level of detail of data collected is appropriate,
- precision – the data collected needs to be precise in term of values,
- relevancy – the data needs to be useful for the purposes for which it is gathered, and
- timeliness – data is entered promptly into the system and available within required time frames [110].

Data collection processes in WSN have been actively explored. [111] presents a Denoising AutoEncoder (DAE) trained to compute the data measurement and the data reconstruction matrices from previously sensed data and gathering the data along a data collection tree. The data measurement matrix is utilized to compress the sensed data in each sensor node, and the data reconstruction matrix is utilized to reconstruct the original data in the sink. There are several procedures [112] of evaluating the data collected in order to achieve target requirements in terms of data accuracy and privacy protection, as well as taking into account challenges with

flexibility in a range of healthcare data collection contexts [113]. Machine learning algorithms and data mining depend heavily on the quality of the collected data because flawed data dramatically lowers their efficiency. To be able to draw any more conclusions and apply the data in an official health system, it is crucial to make sure the data collected is accurate and comprehensive.

Data cleaning aims at how to detect and eliminate data errors originated from the initial data. Data cleaning techniques are the main focus of efforts to address the issue of data cleaning in large WSN. The general flow of the data cleaning process is given in Figure 31.

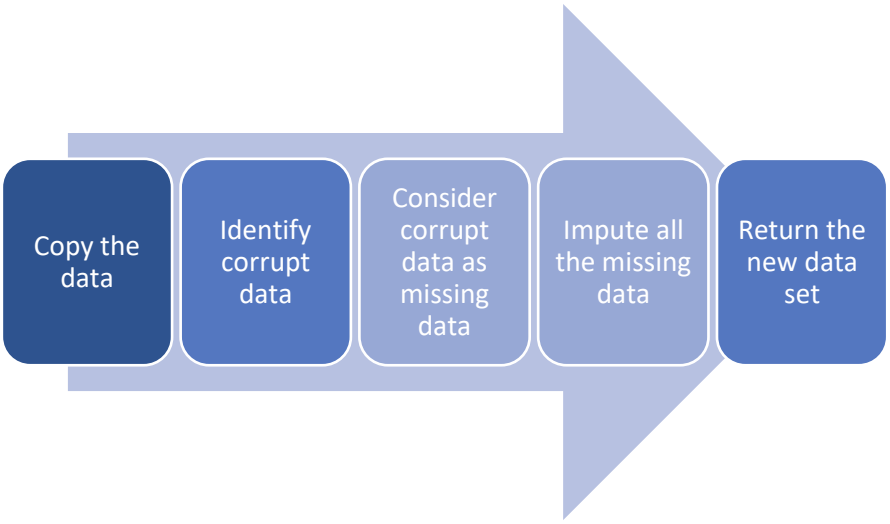

Figure 31. Data cleaning process

The current data cleaning techniques most commonly imply repeated object detection, outlier, value detection, and missing data processing. The impact of missing values in a WSN data set is analyzed [114] by artificially creating missing values in the Intel Lab data set that is publicly available from the Intel Berkeley Research lab. Classification accuracy was then calculated by applying the C4.5 classifier on the original data set as well as the dataset having 10% missing values. The classification accuracy dropped considerably on the data set with the dropped values which suggests that there is a need to clean the erroneous sensor dataset acquired from a WSN.

The traditional approach for data cleansing requires manual review of the data. This is performed by domain experts who manually review the collected data, looking for outliers and unusual events which makes it slow, expensive, tiresome and unscalable. The data cleaning procedure must be automated as a result. An active field of research is automated purification [114]. The accuracy of upcoming readings from the sensor was evaluated by calculating their

probability given the model in the initial techniques, which utilized long-term historical records of a single sensor to construct a probabilistic model of the sensor's behavior over time. The evaluation determines whether the sensor's condition is good or poor according to the evaluation [115]. In environment monitoring WSNs, in which each sensor was encircled by several of its kind for whom the areas interlaced, machine learning-based systems for data cleaning were used. This redundancy was then utilized to learn the interdependencies among sensors and grant accurate predictions without needing long-term historical records [116]. Other researchers propose algorithm solutions for the fault detection and classification in sensor data. In [117], four indicators of the data quality assessment are provided: amount of data, correctness, completeness, and data variation as well as detailed measurement for interrelationships among them. The proximity of the observed value to the true value is indicated by the correctness. The data is deemed accurate if the difference between the recorded value and the actual value of the environment is lesser than a predetermined threshold. The degree of data loss issues in a data set is indicated by its completeness. The raw data volume compared to the necessary data volume is typically used to calculate it. Some measurements, such as displacement, have high volatility when they fluctuate often, but temperature and humidity have low volatility. Volatility is typically used to represent data variance, and it can be evaluated by the acceptable time frame over which the data holds relevant. The study involved uses the data cleaning procedure described above, in which inaccurate data are first identified, then assumed to be missing, and finally all missing values are imputed. A corrupted data detection method known as [118] is used to identify inaccurate data. This method uses the co-appearance matrices, which show the frequency with which each pair of input variables occurs in the data set. By distinguishing the clean records from the records containing noise in the original dataset, CAIRAD creates a clean dataset. FIMUS, an existing missing value imputation method for data pre-processing, is employed to impute the missing data [119]. In order to determine the co-occurrences of the values with other values corresponding to other attributes, FIMUS extrapolates the numerical data points into a number of categories. The capabilities of medical WSNs comprise extremely essential systems, thus it goes without saying that they must be reliable and resistant to sensor failures. To prevent false alarms, it is crucial to identify anomalous data that differ from previous observations and to differentiate between sensor malfunctions and emergency situations. As a result, [120] suggests an algorithm for anomalies identification for WSN in the medical industry. The suggested method initially divides occurrences of detected patient attributes into normal and abnormal categories. Hardware errors, faulty sensors, energy depletion, calibration issues, electromagnetic interference, broken connectivity, compromised

sensors, heart attacks, deteriorating health, etc. can all lead to abnormal readings. Regression prediction is used to differentiate between an erroneous sensor reading and a patient entering a critical state once a very anomalous instance has been identified. [121] demonstrates benchmark datasets for classifying and detecting faults in sensor data. A recent study [122] provides an example of data preprocessing for wearable devices, using adaptive spectrum noise cancellation (ASNC) to remove motion artifacts from photoplethysmography (PPG) signal measured by an optical biosensor and produce clean PPG waveforms for heart rate calculation. Additionally, this cancels out movement-related noise that occurs naturally while a user is moving, which happens frequently when the motion frequency is extremely close to the target cardiac rate. The suggested method does this by making use of the inbuilt accelerometer and gyroscope sensors to dynamically detect and eliminate the abnormalities and achieve accurate heartbeat rate measurement while moving. Instances of pertinent data-driven imputations are provided in [123], where manually collected gathered patient data (BMI) was imputed using various data-driven algorithms, and in [124], where soft sensor data for calculating solar radiation was imputed.

4.1. Data-driven models for cleaning eHealth sensor data

Finding the high-performing prediction model is necessary since using prediction models in medical care demands great precision [123]. Following similar research that was already covered in the previous chapter, numerous models are taken into consideration, including support vector machines, decision trees, random forests, and multiple linear regression.

A. Multiple linear regression

Multiple linear regression (MLR) uses several explanatory variables to predict the outcome of a response variable (Figure 32). Its objective is to represent the linear relationship between the explanatory (independent) variables and response (dependent) variables. The formula for MLR is:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon$$

where y_i is the dependent variable, x_i are explanatory variables, β_0 is y-intercept (constant term), β_p are slope coefficients for each independent variable and ϵ is the model's error term (residuals). Linear regression can only be used when the data has at least two continuous variables, one of which is a dependent variable. The explanatory variables are the parameters

that are used to calculate the dependent variable. The slope coefficients can be used to interpret the outcomes of multiple linear regression while holding all other variables constant. Linear regression works well for both regression [125] and classification problems [126].

The MLR is based on the several conjectures:

- A linear relationship between the dependent variable and the explanatory variable(s) exists
- The independent variables don't have a high correlation among each other
- y_i observations are chosen in an independent and random manner from the population
- Residuals need to have normal distribution with a mean of 0 and variance σ

The coefficient of determination (R^2) is used to measure the amount of the variation in outcome that may be explained by the variation in the explanatory variables. Value of R^2 is between 0 and 1, where 0 means none of the explanatory variables can predict the result, whereas 1 indicates that the result can be predicted with no error from the explanatory variables. R^2 always increases as the additional predictors are incorporated into to the MLR model, whether those predictors are related to the outcome variable or not. Thus, R^2 must not be the only identifier of which predictors should include into and which excluded out of the model.

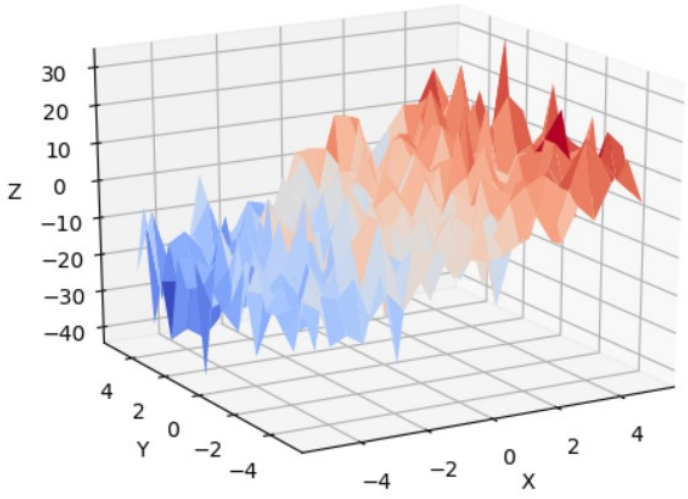

Figure 32. Multiple linear regression in a 3D plane

B. Decision tree

Decision tree (Figure 33) is a prediction model that makes conclusions about the variable's target value from previous observations of the item [127]. Branches represent observations and

leaves represent the conclusions. Tree models in which the target variable has a value which is in a discrete set of values are called classification trees [128]. In such tree structure, class labels are represented by leaves, and the feature conjunctions that result in those class labels are represented by branches. Regression tree is a decision tree in which the outcome variable may assume continuous values (usually a real number).

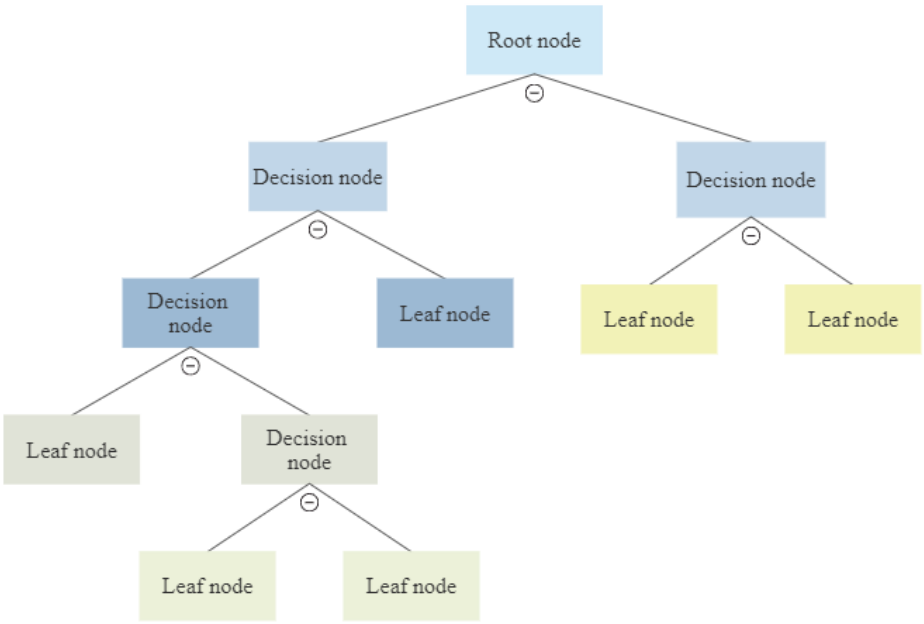

Figure 33. Example of a decision tree

General advantages of decision tree include:

- Easy to visualize and understandable.
- Does not require extensive data preparation.
- The cost is logarithmic in the number of data points in training the tree.
- Can handle data numerically or per category.
- White box model allows for results to be easier to interpret as the conditions are defined by boolean logic.
- Can be validated using statistical tests.

Disadvantages however are:

- Overfitting, i.e., creation of overly complicated trees that leads to data not being well generalized.
- Small data variations can lead to vastly different trees being created.

- Optimal decision tree is known to be NP-complete.
- Some logical concepts are not expressed easily, e.g., XOR or parity.
- Data set needs to be balanced in order to avoid bias.

C. Random forest

As it was already mentioned, one of the disadvantages of decision trees is that they can be unstable as any variation in data may lead to a completely different tree getting generated. In order to mitigate this, an ensemble of trees is used instead. More distinct decision trees that operate as an ensemble are used to create the random forest (Figure 34). In a random forest, each individual tree makes a prediction, and the value that receives the most votes is considered the model's prediction. The trees of the forest and, crucially, their predictions must not be correlated (or need to have low correlations among each other). Generally, a high number of models (trees) with minimal to no correlation operating as an ensemble will perform better than any of the individual models. However, this is more suitable for classification [129] and not regression problems [130].

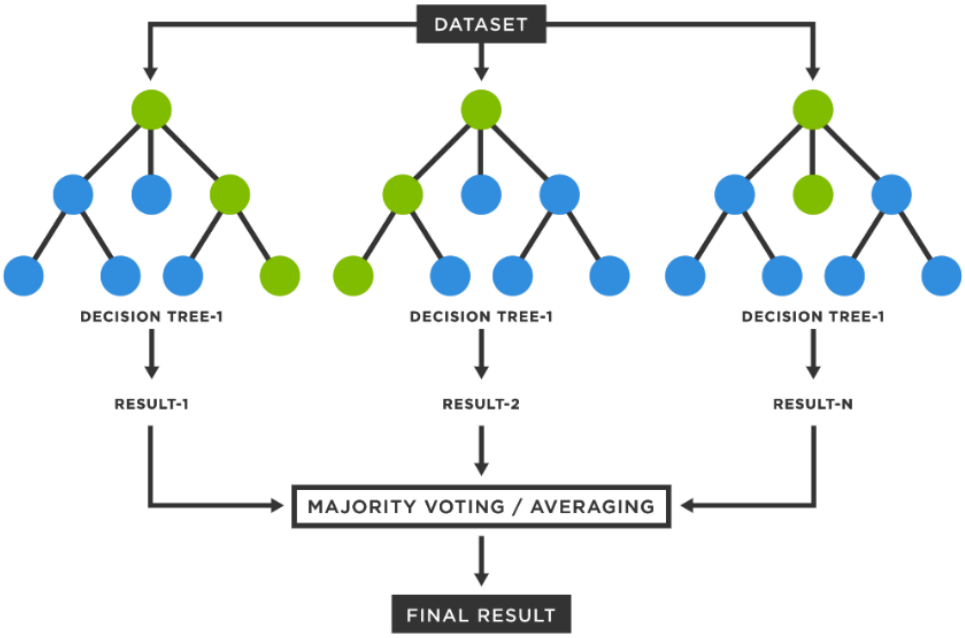

Figure 34. Random forest

D. Neural network

Neural networks (Figure 35) are a set of algorithms which are designed to recognize patterns, modeled loosely after the human brain. All real-world data, including images, sounds, texts,

and time series, must be converted into vectors in order for them to recognize the patterns, which are numerical and contained therein. Neural networks consist of multiple hidden layers that are made of nodes (Figure 36). A node assigns relevance to inputs in relation to the job the algorithm is attempting to learn by combining input from the data with a set of coefficients, or weights, that either amplify or diminish that input. Starting with an initial input layer that receives data, each layer's output serves as the next layer's input at the same time. The significance of the input data is assigned by matching the changeable coefficients of the model with those features. Imputing data using the neural network has been increasing in popularity; one example of using a neural network to impute the missing data for medical IoT applications is given in [131].

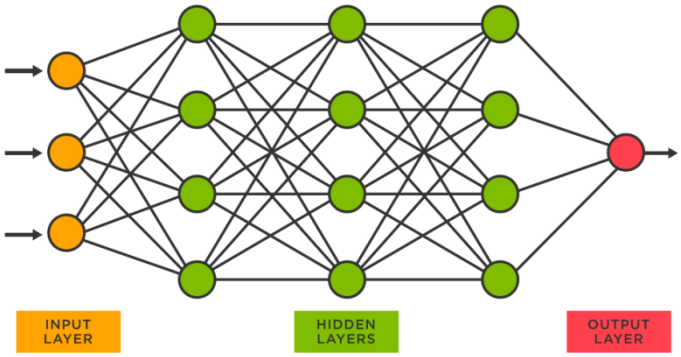

Figure 35. Neural network

Neural networks can learn using human-generated labels, i.e., supervised learning or learn without them by detecting similarities and clustering, i.e., unsupervised learning. Neural networks pose an alternative to standard techniques as they self-train efficiently even when the relationships of variables are non-linear or dynamic. Thus, neural networks can be used in modeling problems which would have been difficult or impossible to describe otherwise as they can be used to discover structure within unlabeled, unstructured data or raw data. However, a neural network is a black box; autonomous, provided only with architecture, random seed numbers and input data. Neural networks also can take longer to train, especially when having massive amounts of data.

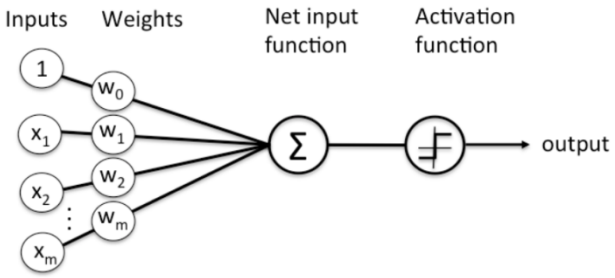

Figure 36. Neural network node

E. Support vector machines

Even though SVM models are usually used for classification objectives, they can be used for both regression [132] and classification [133] tasks. With less processing power, it provides accuracy that is apparent. Finding a hyperplane in an N-dimensional space (N is the number of features) that clearly categorizes the data points is the goal of the support vector machine algorithm (Figure 37). There are a variety of different hyperplanes that might be used to split the two classes of data points. The goal is to identify a plane with the largest margin, or the greatest separation between data points from both classes. Maximizing the margin distance adds some support, increasing the confidence with which future data points can be categorized. A function used in SVM for helping with problem solving is called kernel. Kernel function can be:

- linear: (x, x')
- polynomial: $(\gamma(x, x') + r)^d$, where *d* is specified by parameter *degree*, *r* by *coef0*.
- rbf: $exp\ (-\gamma||x - x'||^2)$, where γ is specified by parameter *gamma*, greater than zero.
- sigmoid: $tanh\ (\gamma(x, x') + r)$, where *r* is specified by *coef0*.

Additionally, custom kernels can be defined.

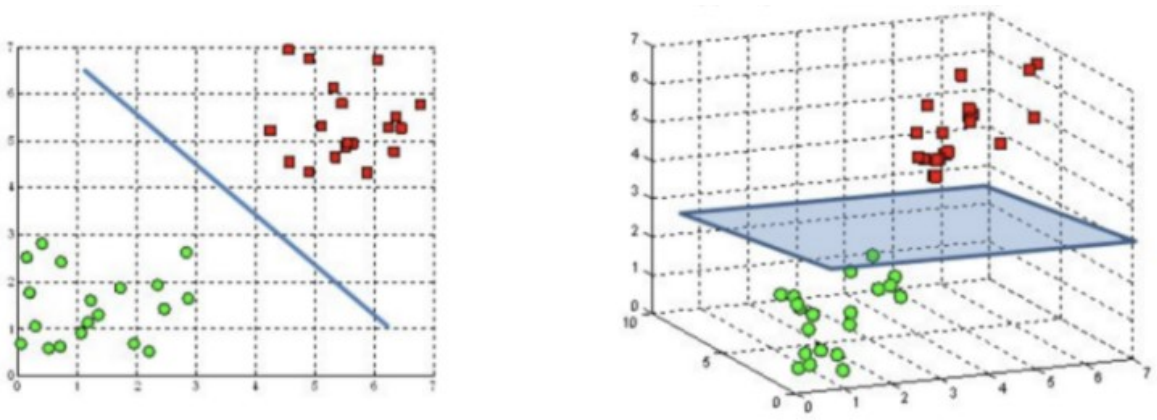

Figure 37. Hyperplanes in 2D and 3D space

Support vector machines are:

- Efficient in high-dimensional spaces, even when the number of dimensions is greater than the number of samples.
- Memory efficient; using support vectors in the decision function.
- Multifaceted as various kernels can be used.

However, compute and storage requirements increase significantly with the number of training vectors.

4.2. Use case on ECG signal: comparison of models

The dataset used is open access MHEALTH dataset [134][135]. The MHEALTH (Mobile HEALTH) dataset comprises body motion and vital signs recordings for ten volunteers of diverse profiles (sex, age and different levels of physical fitness) while engaging in a variety of physical activities. The participant's chest, right wrist, and left ankle were fitted with sensors that record the acceleration, rate of turn, and magnetic field orientation of various body parts. The sensor, which is placed on the chest, generates 2-lead ECG readings, which may be used for routine cardiac monitoring, screening for different arrhythmias, or examining how exercise affects the ECG. A sample rate of 50 Hz is used for all sensory modalities, which is regarded to be enough for capturing human activity. This dataset has been found to generalize to common daily activities. (e.g., running vs. standing still). The example sample in Table 5 includes acceleration from the chest, left ankle and right wrist sensors, as well as a 2-lead electrocardiogram signal, with all sensor measurements (columns) provided. Column Label indicates the action taking place at the time (e.g., activities like standing still, walking and running are all denoted with different numbers). The samples in Table 5 were all collected when the individual was still, therefore they are all identified by the same number (one). In one of the leads' data, 10% of the data was discarded, and the missing values were computed using statistical and machine learning techniques in order to compare the quality of computation of various data-driven models. The estimated values are then contrasted with the actual values, and models are assessed by contrasting the calculated errors.

Table 5. Sample of all measurements readings (columns) from the MHEALTH (Mobile HEALTH) dataset

Acceleration from the chest sensor (X axis) [m/s^2]	Acceleration from the chest sensor (Y axis) [m/s^2]	Acceleration from the chest sensor (Z axis) [m/s^2]	Electrocardio-gram signal (lead 1) [mV]
-9.5987	-1.3757	-1.3238	-0.41863
-9.549	-1.3385	-0.95204	-0.21769
-9.3709	-1.1514	-1.3036	-0.18838
Electrocardio-gram signal (lead 2) [mV]	Acceleration from the left-ankle sensor (X axis) [m/s^2]	Acceleration from the left-ankle sensor (Y axis) [m/s^2]	Acceleration from the left-ankle sensor (Z axis) [m/s^2]
-0.22606	0.99362	-9.7838	1.8228
-0.15908	0.76724	-9.6756	1.7426
-0.15489	0.71657	-9.7453	1.7074
Gyro from the left-ankle sensor (X axis) [deg/s]	Gyro from the left-ankle sensor (Y axis) [deg/s]	Gyro from the left-ankle sensor (Z axis) [deg/s]	Magnetometer from the left-ankle sensor (X axis)
0.28757	-0.61914	0.63065	0.36139
0.28757	-0.61914	0.63065	0.017437
0.29499	-0.60788	0.6169	0.555
Magnetometer from the left-ankle sensor (Y axis)	Magnetometer from the left-ankle sensor (Z axis)	Acceleration from the right-wrist sensor (X axis) [m/s^2]	Acceleration from the right-wrist sensor (Y axis) [m/s^2]
0.54684	-0.012696	-1.5735	-9.341
0.3668	-0.58122	-1.4181	-9.5123
0.54491	-0.44945	-1.4724	-9.3015
Acceleration from the right-wrist sensor (Z axis) [m/s^2]	Gyro from the right-wrist sensor (X axis) [deg/s]	Gyro from the right-wrist sensor (Y axis) [deg/s]	Gyro from the right-wrist sensor (Z axis) [deg/s]
2.8975	-0.070588	-0.64066	0.96552
2.7744	-0.068627	-0.64066	0.97198
2.8474	-0.068627	-0.64066	0.97198
Magnetometer from the right-wrist sensor (X axis)	Magnetometer from the right-wrist sensor (Y axis)	Magnetometer from the right-wrist sensor (Z axis)	Label
0.17578	-0.37549	-1.0758	1
0.17763	-0.19043	-0.71656	1
0.18126	0.17247	-0.72012	1

Scatterplots show the values for a set of data for two variables, usually. A point's position on the horizontal axis is determined by the value of one variable, while its position on the vertical axis is determined by the value of a second variable. The scatterplots for the second lead's original (blue) and computed (green) values for each of the five imputation techniques are shown in Figure 38. Root mean square error (RMSE) and relative root mean square error (RRMSE), which are determined by dividing RMSE by range, serve as the accuracy criteria (max-min).):

$$RMSE_{f_o} = \sqrt{\left[\frac{\sum_{i=1}^{N} \left(z_{p_i} - z_{o_i}\right)^2}{N}\right]}$$

where z_{p_i} is the predicted value for the i[th] observation in the dataset, and z_{o_i} observed value for the i[th] observation in the dataset, and N being the sample size.

Pseudocode for the process of data cleaning and calculating RMSE and RRMSE is as follows:

```
1. Copy the data into a new dataset, called "cleaned_data"
2. Mark randomly 10% of data as erroneous
   Let n be the total number of data points
   Let m be the number of data points to be marked as erroneous, i.e., m = 0.1*n
   For i = 1 to m:
      Choose a random index j between 1 and n
      Set cleaned_data[j] = NaN (i.e., mark it as missing data)
3. Impute missing data using various algorithms
   Let X be the set of features (i.e., sensor readings)
   Let y be the target variable (i.e., the value to be predicted)
   For each feature x in X:
      For each algorithm A (MLR, Decision tree, Random forest, Neural network, SVM):
         Let X_A be the set of features other than x
         Let X_imputed be the imputed version of X using algorithm A on x
         Let X_combined be the combined dataset of X_A and X_imputed
         Fit A on X_combined and y to get a model M_A
         Predict the missing values in x using M_A and X_combined
         Set the missing values in cleaned_data[x] to the predicted values
4. Return cleaned_data
5. Compare cleaned_data to original data set and calculate RMSE and RRMSE
```

Step-by-step explanation of each part of the pseudocode:

First step simply creates a new dataset called "cleaned_data" that is a copy of the original dataset. This is done to ensure that the original dataset is not modified during the cleaning and imputation process. Second step randomly selects 10% of the data points in the dataset and marks them as erroneous by setting their values to NaN, which represents missing data. This is done to simulate missing data that may occur in real-world datasets. Third step involves using various algorithms to impute the missing data in the dataset. For each feature in the dataset, the algorithm first creates a version of the dataset where the missing values for that feature have been imputed using the chosen algorithm. It then combines this imputed dataset with the original dataset (with the missing values still present) and fits a model using the chosen algorithm on this combined dataset to predict the missing values. The predicted values are then used to replace the missing values in the cleaned dataset. Fourth step simply returns the cleaned dataset that was created in step 1 after the missing data has been imputed using the various algorithms. Final fifth step involves comparing the cleaned dataset to the original dataset using two previously mentioned metrics: RMSE and RRMSE. The purpose of this step is to evaluate how well the imputation algorithms performed in filling in the missing data.

The errors vary between subjects (Table 6), and multiple linear regression and neural networks consistently outperformed all other methods. As a result, it was decided to use those techniques to further refine the model in the chapter that follows.

Table 6. Root mean square error (RMSE) and relative root mean square error (RRMSE) for predictions of ECG signal using various models for imputation of missing values for three different participants

	RMSE (person A)	RRMSE (person A)	RMSE (person B)	RRMSE (person B)	RMSE (person C)	RRMSE (person C)
Multiple linear regression	0.27320	0.04362	0.21398	0.03184	0.05657	0.01366
Decision Tree	0.27678	0.04419	0.24916	0.03708	0.10049	0.02427
Random Forest	0.27606	0.04408	0.21263	0.03164	0.05698	0.01376
Neural Network	0.27161	0.04337	0.21189	0.03153	0.05660	0.01367
Support Vector Machines	0.28006	0.04472	0.21446	0.03198	0.05658	0.01366

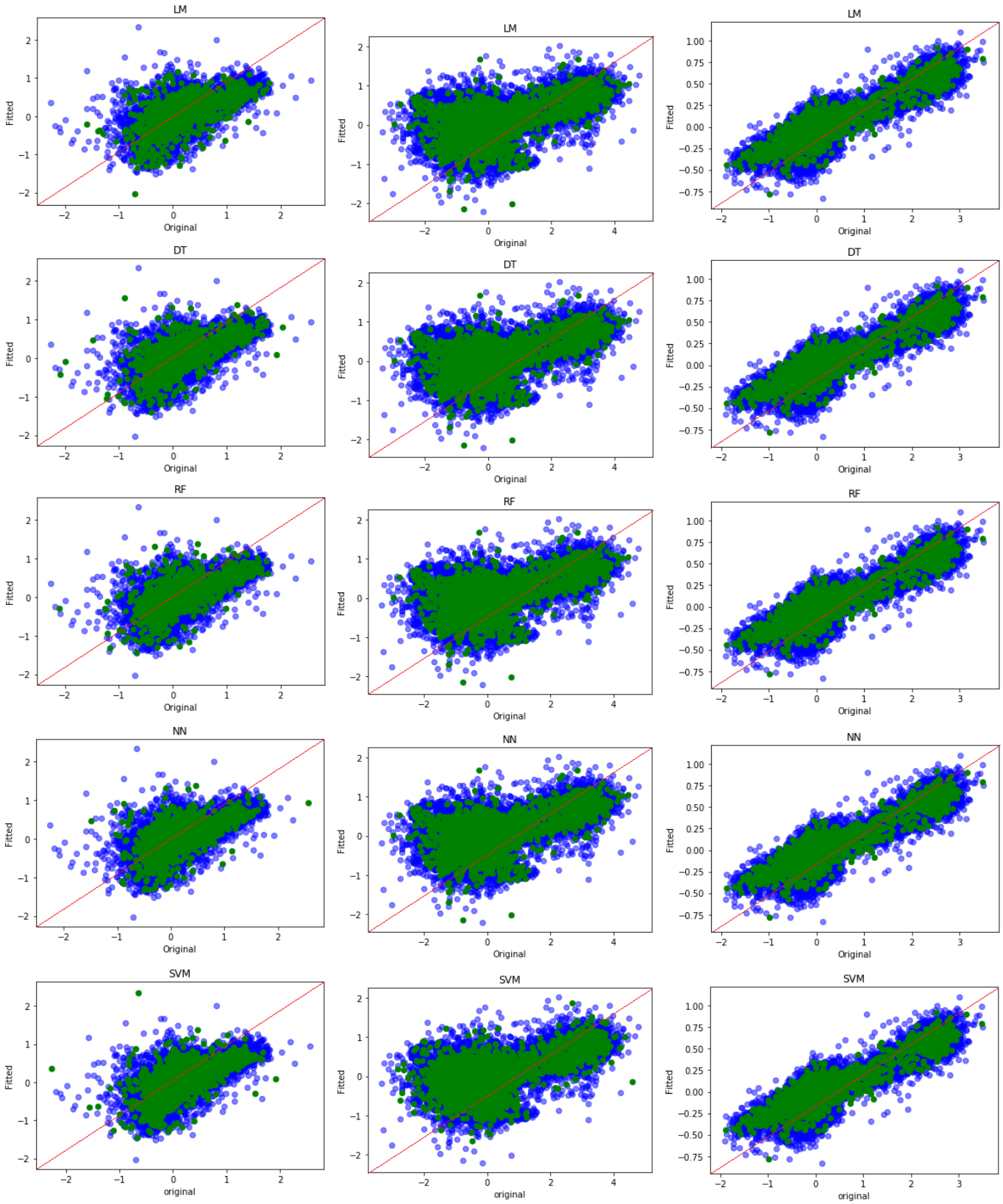

Figure 38. Scatterplots for original (blue) and computed (green) values of the second lead for five imputation methods (rows); LM – linear model with multiple imputation, DT – decision tree, RF – random forest, NN – neural network, SVM – support vector machine for three different subjects (columns). Blue – original values, green – missing imputed values

4.3. Model improvements: classification algorithms

The method [136] employs decision trees and forests to discover segments of a data set where the records within a segment have higher similarity and attribute correlations in order to improve imputation outcomes. The missing data are then imputed using the correlation and similarity. This method can be used for a wide range of data, and as a result, it has numerous potential uses. Nine publicly accessible data sets of various information, such as housing or credit approval, were experimentally examined, and the findings showed an apparent superiority of such statistically-based methodologies. Assuming that there is a higher correlation in the portions that are separated by the physical activities engaged in (such as sitting motionless, walking, or running), which reflects the established relationship between physical activity and heart activity [137]. This data is entered by hand into the MHEALTH dataset by the monitor (last column Label). However, if very accurate data segment categorization is feasible, it might be used to any comparable m-health dataset. Additionally, as a lot of health data originates from smart wristbands or smartwatches, such as [138], the calculations that follow concentrate on the data from the sensors that are located on the right wrist, both gyro and acceleration (columns 15-20). The final step is to divide the dataset into parts based on the activities categorized in the previous stage. Walking and other light motions are considered light activity, whereas sitting and standing motionless are considered inactive. Jogging, running or exercising is a medium activity. The imputation is then carried out once more, this time with multiple linear regression and neural networks because they produced the greatest outcomes in the previous chapter. Pseudocode:

```
1. Copy the data into a new dataset, called "cleaned_data"
2. Mark randomly 10% of data as erroneous
   Let n be the total number of data points
   Let m be the number of data points to be marked as erroneous, i.e., m = 0.1*n
   For i = 1 to m:
      Choose a random index j between 1 and n
      Set cleaned_data[j] = NaN (i.e., mark it as missing data)
3. Impute missing data using MLR / Neural network taking the new variable created
in previous step as additional input
    for column in new_data:
        missing_indices = new_data[column].isnull()
```

```
        new_data.loc[missing_indices,                    column]                    =
impute(new_data[~missing_indices][column], activity_levels[~missing_indices])
4. Return cleaned data
    return new_data
5. Compare cleaned data to original data set and calculate RMSE, RRMSE
```

First step creates a new dataset called cleaned_data that will be used to store the cleaned sensor health data. The original dataset is not modified. Second step randomly selects 10% of the data points in the dataset and marks them as erroneous by setting their value to NaN, which indicates missing data. This is done to simulate the presence of errors or missing data in real-world sensor data. Third step uses a machine learning algorithm, either multiple linear regression (MLR) or a neural network, to impute missing data in the dataset. The algorithm takes the cleaned_data dataset with the randomly marked missing values as input, along with an additional variable that was created in the previous step to indicate which data points were marked as erroneous. The algorithm then imputes the missing values in the dataset using the other available sensor data. Fourth step returns the cleaned dataset with imputed missing values. The new_data dataset will now be free of any missing values or errors that were previously present. Final step involves comparing the cleaned dataset with the original dataset to evaluate the accuracy of the cleaning process. RMSE and RRMSE are calculated to quantify the difference between the two datasets. This step provides a measure of how much the cleaning process changed the original dataset and how well the imputation algorithm performed in filling in the missing data. By randomly marking some data as erroneous and then using a machine learning algorithm to impute missing values, the cleaned dataset can be compared to the original dataset to evaluate the effectiveness of the cleaning process. The goal was to optimize the data cleaning model and improve accuracy by using an extra variable - intensity of physical activity, which is known to correlate with the values being imputed. These imputed entries are then combined into a complete dataset that has no missing information. For combined datasets, values for RMSE and RRMSE are shown in Table 7. While the results are consistent across categories, the accuracy increased by a total of 10% to 17%.

Table 7. Root mean square error (RMSE) and relative root mean square error (RRMSE) for predictions of full ECG signal using multiple linear regression and neural networks method for imputation of missing values by activity for three different participants

	RMSE	RRMSE	Improved accuracy (%)	Person
Multiple linear regression	0.23764	0.03794	13,2	Person A
Neural Network	0.22992	0.03671	15,34	
Multiple linear regression	0.18340	0.02728	14,29	Person B
Neural Network	0.17541	0.02610	17,21	
Multiple linear regression	0.05098	0.01231	9,88	Person C
Neural Network	0.05094	0.01230	10,0	

4.4. Discussion

It is crucial for the suggested strategy to have an effective classifier that can observe and classify various physical activities. In [139] and [140], a single waist-worn accelerometer is used to identify physical activities, producing robust findings in a lab setting. Results from the free-living scenario, however, revealed noticeably worse performance. Three data mining algorithms (Decision Tree, Random Forest, and PART algorithm) were tested for recognition of 33 different physical activities in [141], with an overall accuracy of 96.52% while utilizing 6 sensors. Similar results were obtained by [142] using a multi-sensor system using a 3D accelerometer and gyroscope to attain 97.38% accuracy. In recognition of physical activity using smartphone sensors, [143] achieves an accuracy of 88% while [144] classifies five common activities with 98% accuracy. This is due to the integration of acceleration sensors and gyroscopes in smartphones. The results are significantly influenced by the position of the smartphone, even though employing smartphone sensors is less intrusive than multi-sensor setup. Long processing delays also interfere with other typical mobile phone tasks. Body-

mounted smartwatches make it possible to monitor activities unobtrusively. In addition, the precise location aids activity recognition research by relieving the strain of determining the location of the device. Ultimately, the results offer a satisfactory level of classification accuracy and a show promise for their application in imputing sensor data. Some of the problems include performance inconsistencies between the lab and free-living environments, variances in somatotype and athletic habits, which have a significant impact on the imputation and classification outcomes.

5. Specification of semantic data constraints and validation of data

After ensuring quality of data, it is necessary to organize and reformat the data collected across various devices into certain, predetermined format for simple integration into the EHR. This also implies that all data and information exchange must carefully adhere to the most recent norms and regulations [145]. The following chapter will provide more detail on this.

5.1. Materials and methods

The study involved three steps: 1) data collection, 2) modeling the process of parsing, data verification and validation, and 3) verifying the proposed process through a use-case study employing datasets including various pertinent data types.

5.1.1. Data collection

Two datasets have been utilized to validate the validation process outlined in the next chapter. The Fitbit Versa 2 smartwatch wristband was used by sixteen people to report their activities across a five-month period for the PMData Dataset [146], which was made available via Simula Open Datasets. Additionally, during the period of two months, the OxyBeat [147] dataset, which includes heart rate, body temperature, and oxygen saturation (SpO2), was gathered for the purpose of this study using Fitbit Versa 3. This was done to ensure a more robust use-case scenario by adding a number of additional data types, with an emphasis on COVID-relevant data types. When compared to the total amount of hemoglobin in the blood, SpO2 indicates the proportion of oxygenated hemoglobin (hemoglobin containing oxygen) (oxygenated and non-oxygenated hemoglobin). Deoxygenated blood has a darker red hue than fully oxygenated blood in the arteries and arterioles and travels back to the lungs through veins. Even during physical activity and sleep, blood oxygen saturation (SpO2) tends to fluctuate very little; during the day, blood oxygen levels typically range from 95 to 100%. Since less air is breathed in total while a person sleeps, their SpO2 will often be lower than it is during the day. SpO2 levels at night are often >90%. Additionally, the Fitbit Versa 3 smartwatch wristband was used to collect the data as it offered access to readings for oxygen saturation. One of the most well-known commercial wearable activity trackers, Fitbit, has sold over 105 million units globally since 2010 and has close to 30 million active users. The Health Metrics Dashboard (Figure 40), which

tracks parameters including breathing rate, heartbeat rate variability, and SpO2 — all crucial metrics when it comes to disease diagnosis — was incorporated into the devices in the third quarter of 2020 [148]. Wearables' triaxial accelerometers record spatial body movements. Utilizing specialized algorithms, motion data is evaluated. In order to measure health-related metrics like steps taken, exercise time, or sleep time, patterns of motion are discovered. Wearables were first designed as a consumer product to encourage people to exercise and be physically active, but their use as research equipment and a patient support aid is growing [149]. 260 clinical experiments have been registered at ClinicalTrials.gov since 2011 that employed Fitbit exclusively for data collection [150]. Steps were the main outcome of interest for the majority of the studies in question, followed by time spent exercising or sleeping, heart rate, and energy expenditure. Smart wearables, particularly those worn on the wrist, have demonstrated that they are dependable, durable, and acceptable [151].

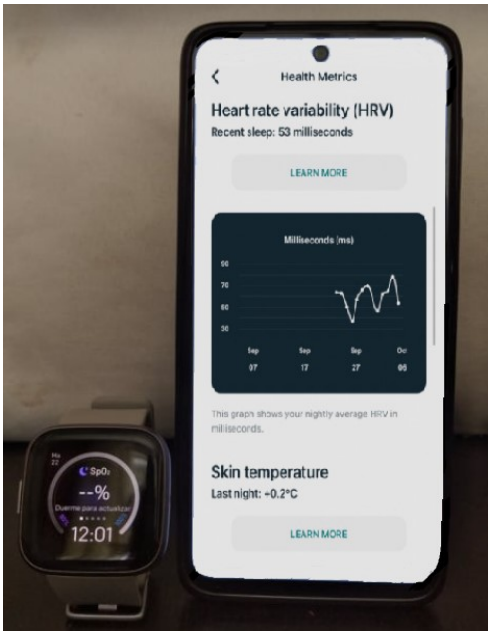

Figure 40. Health Metrics Dashboard

Heart rate, oxygen saturation (SpO2), and body temperature are the health-related indicators taken into account in this study. The pulse rate is one of the critical indications that all personal trackers monitor. Additionally, the module for cleaning sensor data of this specific datatype has already been created and extensively discussed in earlier published paper [90]. As a result, this will also serve as the model's initial syntax and semantic validation example. The HL7 Structure Definition of HeartRate [152] must be adhered to in order to comply with HL7 standards. The

structure is developed from vital indicators observed during observation [153]. The tracker exports heart rate information as a JSON file, as seen below.

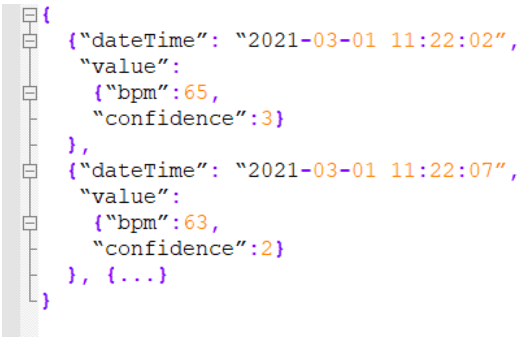

Figure 41. An example of Heartbeat rate data (JSON)

5.1.2. Process for parsing, verifying and validating the data

Five seconds pass between heart rate observations, giving the PMData dataset a total of more than 1.5 million readings for the specified time frame. Each reading includes a timestamp, a "bpm" field that indicates the heart rate value, or beats per minute, and the reading's confidence level. The device's level of confidence in the accuracy of the heartbeat measures varies from 0 to 3, with 3 denoting the highest level of confidence. For cleansing eHealth data, a data-driven methodology was used in earlier research [90]. It has demonstrated an accuracy improvement of between 10 and 17% and imputes inaccurate data using neural network methods. Copying the data, identifying corrupt data, treating corrupt data as missing data, imputing all missing data, and then producing the new dataset are all steps in the data cleaning process. In this instance, the data points with confidence equal to 0 were those chosen for imputation. As a result, cleaned data includes imputed values for data points with confidence level zero as well as original values for those with confidence levels 1-3. Then, using Schematron-based validation, clean data is parsed and examined. This process is illustrated in Figure 42.

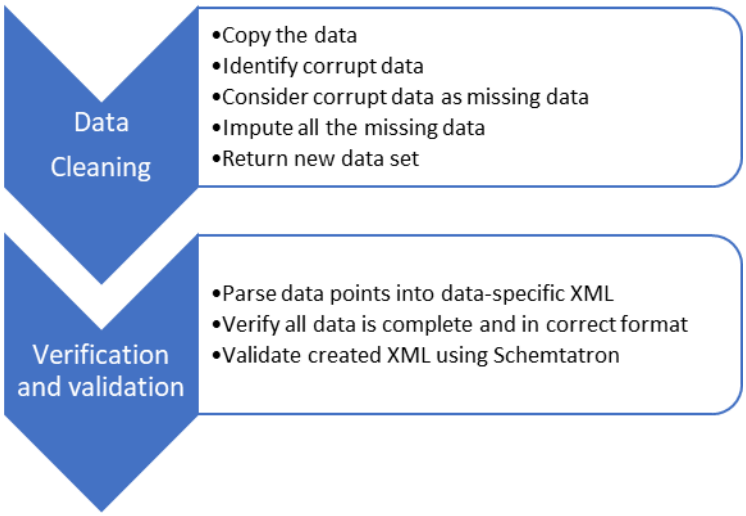

Figure 42. Process of data cleaning and validation

Large amounts of data must be supported by the parser effectively. Additionally, it must be reliable and not crash. A JSON parser that is SIMD-accelerated (single instruction, multiple data) and can process data at speeds of up to 2.2 GB/s was designed and implemented in Python using the pysimdjson module. It has two stages, with translator components in between: the first stage processes input in 64-byte batches, while the second stage creates a "tape representation" after it has been translated. The technique is described in detail in Figure 43 below.

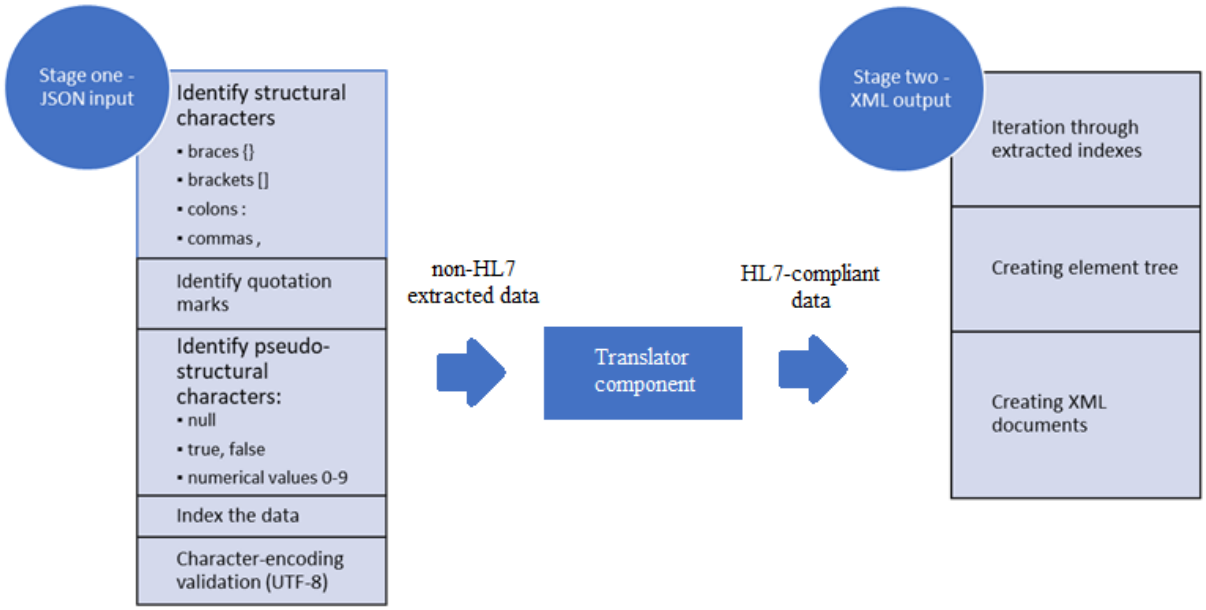

Figure 43. Process of data parsing

Pseudocode for data parsing process, including the algorithm used to extract and convert data within the Translator component is given below:

```
1. Identify structural characters (braces, brackets, colons, commas)
2. Identify pseudo structural characters (null, true, false, numerical value)
3. Index the data
Let data be an empty array
For each character in JSON_data:
    If the character is a brace or bracket, push it onto the data array
    If the character is a colon or comma, push it onto the data array
    If the character is a quotation_mark, push the entire string within the quotes
onto the data array
    If the character is a pseudo structural character or numerical value, push it
onto the data array
4. Validate UTF-8 encoding
5. Translate data into HL7 compliant format
Extract non HL7 data and convert to HL7-compliant data
       Input: list of indexed data lines from JSON (L) that need to be translated
       Input: name ID to verify type of Resource (R)
       Output: data mapped to correct HL7 Resource element (M)
Use name ID to identify Resource type (R) using keywords from Resource IDs (K)
var: current line (l) to convert
loop for each l ϵ L:
       var: list of elements (E)
       split(s) into E
       tokenize(s) into E
       var e ϵ E is current element
       loop for each e ϵ E:
              var current map element m(K,e)
              correctFormat(e)
              tag(m(K,e))
              add to output map M(k,e)
       done
done
6. Creating element tree
Let root be a new XML element called "health_data"
For each item in HL7_data:
    Let element be a new XML element with the name of the HL7 element and the value
of the HL7 element
```

```
    Append the element to the root element
7. Create XML document
    Let xml_doc be a new XML document
    Append the root element to the xml_doc
    Return xml_doc as a string of XML-formatted data
```

The input parameters for the algorithm are the list of indexed data lines from JSON parser (L) and name ID to identify the Resource (R). Output is a data map of elements of a specific HL7 Resource. Each line is split into tokens. For each token, corresponding element of the resource is found (K). Elements are added into an output map containing values (e) mapped to elements (K) of a particular Resource (R). This is passed onto the XML parser.

Extensible Markup Language (XML) documents are the data format required for EHR. Elements and attributes make up XML documents. The following schemas can be used to specify an XML document's permitted structure:

- Document Type Definition (DTD),
- Microsoft XML-Data Reduced (XDR), or, most commonly used
- XML Schema definition language (XSD).

Schema validation was the only validation method initially used for XML documents. This signifies that an XML document was considered valid if it had satisfied schema validation. While ensuring document structure is correct, schema validation is unable to verify conditional and integrity criteria. The suggested procedure for validating personal healthcare papers (Figure 44) would check the document's structure before validating the content and characteristics of the document. Finally, it is necessary to validate any extra limitations that might additionally exist.

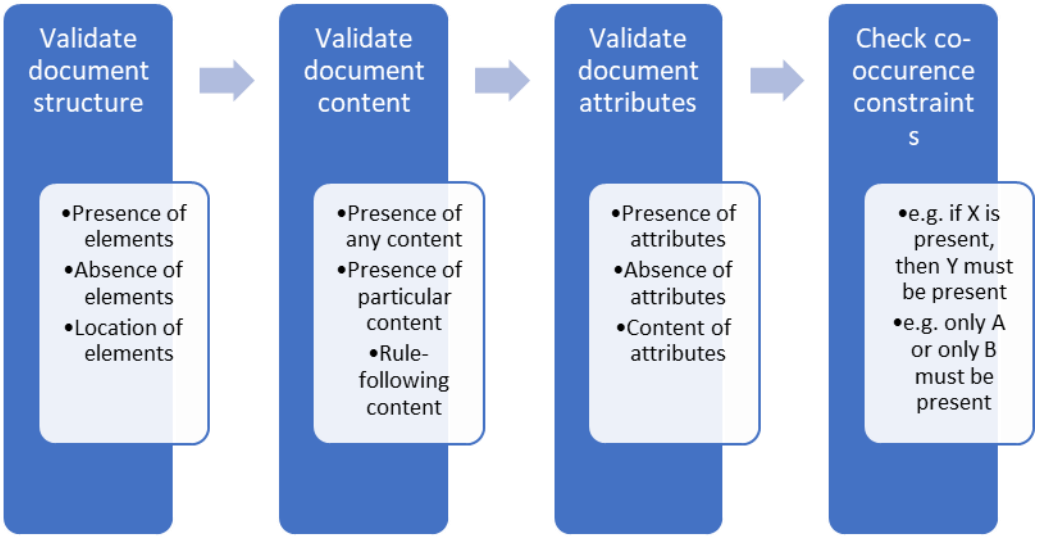

Figure 44. Expected validation process

Schematron is a structural schema validation language based on rules and expressed in Extensible Markup Language (XML). It is frequently used for making assertions about the presence or absence of specific patterns in XML trees and has the ability to create restrictions in a way that other XML schema languages, such as XML Schema and Document Type Definition (DTD), cannot (Figure 45).

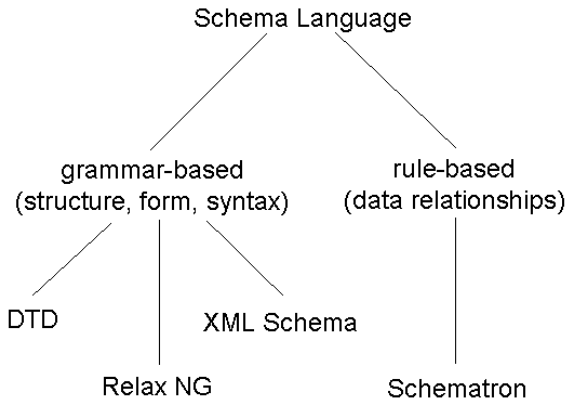

Figure 45. Classification of schema languages

Skeleton implementation of ISO Schematron is a four-stage XSLT pipeline [197]. XSLT (Extensible Stylesheet Language Transformations) is a declarative language used for transforming XML documents into other formats, such as HTML, plain text, or another XML format. XSLT is part of the XSL (Extensible Stylesheet Language) family of languages and is

often used in conjunction with XSL-FO (Extensible Stylesheet Language Formatting Objects) for generating printable documents. XSLT uses XML-based templates and rules to describe how to transform an input XML document into a desired output format. XSLT works by applying templates to the nodes of the input XML document, which match specific patterns defined in the XSLT stylesheet. Each template contains instructions on how to transform the matched node and its children into the desired output format. The first two stages are macro-processors and only necessary if expert features are being used.

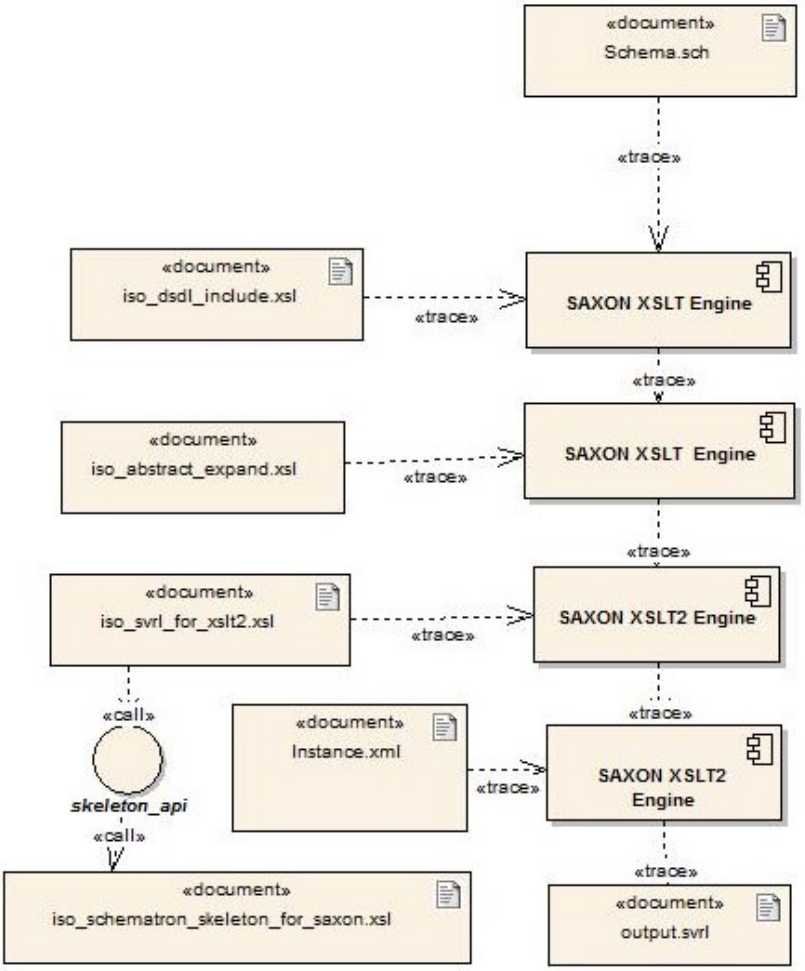

Figure 46. UML diagram showing Schematron integration [198]

Here's how Schematron validation of an XML document works, step by step:

- Schematron schema: Create a Schematron schema, which is an XML document that defines the validation rules for desired XML documents. A Schematron schema typically consists of two parts: a set of patterns that define the rules, and a set of assertions that implement those rules.

- XSLT compiler: XSLT Compiler compiles Schematron schema into an XSLT stylesheet. The XSLT compiler takes Schematron schema and generates an XSLT stylesheet that can be used to validate XML documents against the rules defined in the schema.
- XSLT processor: Then use an XSLT processor, such as Saxon or, in this case, Xalan, to run the XSLT stylesheet generated by the compiler. The XSLT processor takes your XML document and applies the rules defined in the Schematron schema to it.
- XSLT checker: The XSLT processor produces an intermediate result in the form of an XML document that contains information about the validation results. This intermediate document is then processed by an XSLT checker, which is another XSLT stylesheet that converts the intermediate result into a human-readable report that lists the validation errors and warnings.
- Document with XSLT checker: Finally, run XML document through the checker. The checker applies the rules defined in the Schematron schema to your XML document using the XSLT processor and produces a validation report that shows any validation errors and warnings that were found.

In summary, Schematron validation involves creating a Schematron schema, compiling it into an XSLT stylesheet, running the stylesheet against your XML document using an XSLT processor, processing the intermediate result with an XSLT checker, and producing a validation report that shows any errors or warnings. This process is pictured below:

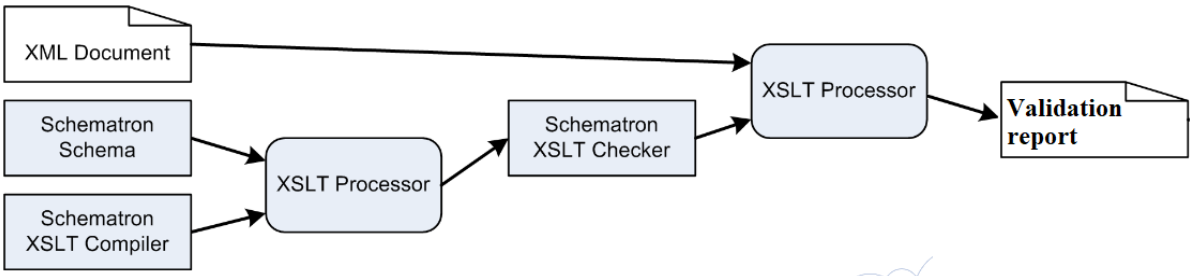

Figure 47. Process of Schematron validation

This requires specific attributes, content control of certain elements by another element or even specification of requirements between multiple XML files [42]. "ISO as Information Technology, Document Schema Definition Languages (DSDL), Part 3: Rule-based validation, Schematron (ISO/IEC 19757-3:2016)" has standardized Schematron. Schematron is the most widely used industry standard language for defining data constraints. To express analogous

health data (e.g., heartbeat rate), different devices follow varying XML syntax. The XML documents must have correct syntax in addition to being meaningful in terms of their semantics for personal health data exchange and integration to be successful. The goal is to simplify things by creating and maintaining a single Schematron document for similar health data across various personal tracking devices. Schematron may construct relationships or constraints where one element depends on another element and can demand the presence of some, all, or none of the attributes in any given element. Complex rules and restrictions required for semantic validation can be implemented with it. Syntax validation refers to grammar-based validation, such as data field type (number, string or Boolean) and simple value constraints (minimum or maximum length of a string, or minimum and maximum values of a number). Semantic validation means constraint rules among the data fields in a document. For example, if data field value 'A' is 'X' then field 'B' must be 'Y'. The data coming from different manufacturers or devices usually means data structure is proprietary to that device (or line of devices) and is described differently for the same specific measure (e.g., heart rate). In order to process such data, data points need to be confined to a single shared schema, regardless of the source.

Schematron makes review of a document according to its context and has a set four-layer hierarchy:

I. Phases
II. Patterns
III. Rules
IV. Assertions / Reports

Assertion is a common element in Schematron schema for defining assertions. It specifies data constraints which will be inspected from the specified context of the XML document. The test attribute that the assert element contains is an XSLT pattern. This is expressed using Schematron assertions as follows:

```
<assert test="@id">The element Observation must have an id attribute.</assert>

<assert test="count(*) = 2 and count(Category) = 1 and count(Code)= 1">The element
Observation must have the child elements Category and Code.</assert>
```

Unless the condition is fulfilled, an exception is raised, with the message displayed being the content of the assertion element. Alternatively, report can be used instead of assert.

The distinction between an assert and a report is that the former produces an error when its assertion is broken, whereas the latter does so when its condition is fulfilled.
The rule element, which has a context property with a value that must match an XPath Expression that chooses at least one node in the document, is used to create Schematron rules. Where the assertion must be applied is specified by the context attribute. The context is fixed to the Observation element in the example above, so the Schematron rule with the Observation element as the context should appear as follows:

```
<rule context="Observation">
      -assert 1-
      -assert 2-
</rule>
```

Pattern element groups various rules together and has a name attribute which is shown to the user when the pattern is applied. For example:

```
<pattern name="Check structure">
      -rule 1-
      -rule 2-
</pattern>
```

Finally, the Phase element allows progressive validation of constraints in stages rather than validating everything at once. Furthermore, it is possible to use Schematron for validation of XML instance documents which use namespaces, with each of the namespaces being declared in the schema prefixed by the ns element. Namespace URI and namespace prefixes found in ns element are defined by its two attributes, uri and prefix. For the XML instance document defined in namespace “http://hl7.org/fhir”, the Schematron schema would look like this:

```
<sch:schema xmlns:sch="http://www.ascc.net/xml/schematron">
    <sch:ns uri="http://hl7.org/fhir" prefix="ex" />
      -pattern 1-
      -pattern 2-
</sch:schema>
```

It is necessary to standardize the personal health data to EHR data transfer in this scenario through the use of schematrons so that it has the proper medical context and structure. Together

with other EHR data (such as past diagnoses, medication usage, doctor visits, and test results), this information can help medical professionals evaluate patients remotely, identify problems early, and provide better healthcare. A project involving industry and healthcare professionals called Integrating the Healthcare Enterprise (IHE) aims to improve healthcare. It offers guidelines, resources, and services for managing and exchanging healthcare data interoperability. "Engage physicians, health authorities, industry, and users to create, test, and deploy standards-based solutions to critical health information needs," is the organization's stated objective [43]. This is accomplished by setting standards, outlining requirements, and giving developers technical frameworks and guidelines to work within. IHE Technical Frameworks [44] outline how to put previously established standards into practice in order to achieve warranted medical information exchange, enable feasible and effective system integration, and provide the best possible patient care. These are regularly expanded and maintained by the IHE Technical Committees on an annual basis, following a period of public review. Among others, Technical Frameworks (TF) described includes those of Anatomic pathology, Dental, Cardiology and IT infrastructure. IT infrastructure describes integration profiles [45] as well as ITI transactions ITI-1 to ITI-28 [46], ITI-29 to ITI-64 [47], ITI-65 and greater [48], together with the metadata [49][50]. Cardiology TF specifically encourages the creation of a variety of implementation profiles, such as cardiac or intravascular imaging, resting ECG (REFW), or stress testing workflow (STRESS), many of which would profit from the aforementioned incorporation of aggregated personal health data (such as ECG information) into a formal EHR [51][52]. Furthermore, Quality, research and public health TF under Supplements for Trial Implementation invites beginning of development for several implementation profiles, among them, Aggregate Data Exchange (ADX) [53].

"Interoperable public health reporting of aggregate health data" is the purpose of ADX. Periodic (weekly, monthly, quarterly, or annual) reports from a health facility to an administrative jurisdiction are the most typical use cases for ADX. Two normative message structure definition files must be created in order for a Content Data Structure Creator to determine the structure of the XML data that will be transmitted between a Content Creator and Content Consumer:

- A Data Structure Definition (DSD) file that follows the standard schematron
- The ISO Schematron schema and the W3C XML Schema Definition (XSD), both of which need to match the output of the normative XSLT transformation from DSD to XSD and from DSD to schematron.

Rules-based schema language is required for the phases in the validation process that deal with contents and constraints. Grammar-only based schemas, like XSD, can't possibly achieve the

desired level of validity. Alternative choices include the XML schema languages Tree Regular Expressions for XML (TREX), REgular Language for XML Next Generation (RELAX NG), and Schematron, which can all check both the structure and the contents of an XML document. In contrast to TREX and RELAX NG, Schematron is able to make assertions about patterns of occurrence anywhere in the text. The ability to nest Schematron schemas inside XSD schemas is another benefit of utilizing Schematron. Due to this, Schematron is the most appropriate schema language for the process of validation described in this study. Using the aforementioned datasets, a schematron-based validation procedure will be defined and subsequently verified.

To express comparable health data, different devices develop varying XML syntax (e.g., heartbeat rate). The XML documents must represent a complete FHIR resource and have acceptable syntax in order for personal health data transfer and integration to be successful. The goal is to simplify things by creating and maintaining a single Schematron document for similar health data across various personal tracking devices. Schematron may construct relationships or constraints where one element depends on another element and can demand the presence of some, all, or none of the attributes in any given element. Complex rules and restrictions required for semantic validation can be implemented with it. The rule element, which has a context property with a value that must match an XPath Expression that specifies one or more nodes in the document, is used to create Schematron rules. The place of application of the assertion is specified by the context attribute. The context is set to the Observation element in the instance above, thus the Schematron rule with the Observation element serving as the context should seem as follows (Figure 48).

```
<rule context="Observation">
    <assert test="@id">The element Observation must have an id attribute.</assert>
<assert test="count(*) = 2 and count(Category) = 1 and count(Code)= 1">
    The element Observation must have the child elements Category and Code.</assert>
</rule>
```

Figure 48. Rule instance for FHIR Resource Observation

Heartbeat rate Schematron for validating data with the intention of being included in EHR must have the rules specified in order (Figure 49) to satisfy all of the aforementioned requirements and adhere to the standards.

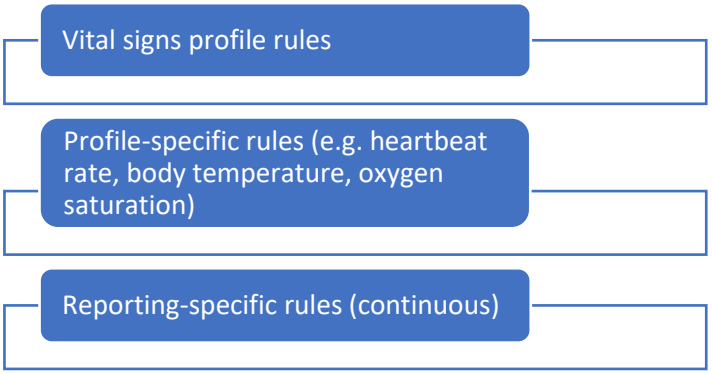

Figure 49. Rules hierarchy for heartbeat rate Schematron

5.1.3. Process verification: use-case study

In a use-case study, the defined method was validated using two sets of data that contained a range of pertinent types of data. The results are shown in the chapter that follows.

5.2. Results

The objective was to investigate the inclusion of personal health-related data collected by mass-market wearables into the central HIS's EHR. Data including cardiac rate, body temperature, and oxygen saturation (SpO2) was collected via use of activity tracker for the goal of this research in order to evaluate the data cleaning and validation process described in this work. The data was cleaned using the data-driven methodology for cleaning of eHealth data [90], which imputes inaccurate data using neural network techniques and has previously improved accuracy by 10-17%. Two-stage parsing model has been designed to test the given validation process in a use-case scenario. Two months' amount of readings for each of the three data types evaluated equates to over 50MB of JSON files with over 2 million data points, which a Ryzen 5 5600X six core processor analyzed in 58 seconds (set to run at 3.7 GHz in 64-bit mode). This has been done for all data points, with a success rate of one hundred percent, in terms of reliability.

As a thorough explanation of the necessary requirements, a detailed overview of all the rules established is provided in the remainder of this chapter. This made it easier to make Schematrons for three different sorts of data. This is applied in the procedure' last stage of data validation.

5.2.1. Heartbeat rate

FHIR HeartRate Structure Definition provides a description of heartbeat rate data. Schematron for heartbeat rate type data is presented in Figure 50.

```
<sch:schema xmlns:sch="http://purl.oclc.org/dsdl/schematron" queryBinding="xslt2">
  <sch:ns prefix="f" uri="http://hl7.org/fhir"/>
  <sch:ns prefix="h" uri="http://www.w3.org/1999/xhtml"/>
  <sch:pattern id="observation">
    <sch:title>Observation</sch:title>
    <sch:rule context="f:Observation">
      <sch:assert test="not(exists(f:dataAbsentReason)) or (not(exists(*[starts-with(local-name(.), 'value')])))">dataAbsentReason
        SHALL only be present if Observation.value[x] is not present (inherited)</sch:assert>
      <sch:assert test="not(exists(f:component/f:code)) or
        count(for $coding in f:code/f:coding return parent::*/f:component/f:code/f:coding[f:code/@value=$coding/f:code/@value
        and f:system/@value=$coding/f:system/@value])=0">Component code SHALL not be same as observation code (inherited)</sch:assert>
      <sch:assert test="f:id/@value = 'heart-rate'">Observation is not heartrate type observation</sch:assert>
      <sch:assert test="exists(f:subject/f:reference/@value)">Patient must exist and be uniquely defined</sch:assert>
      <sch:assert test="matches(f:effectiveDateTime/@value, '^(\d{4}-\d{2}-\d{2}T\d{2}:\d{2}:\d{2}\+\d{2}:\d{2})$')">Observation
        needs to have proper format for dateTime</sch:assert>
      <sch:assert test="exists(f:valueQuantity/f:value/@value)">Observation needs to have value measured</sch:assert>
    </sch:rule>
  </sch:pattern>
  <sch:pattern id="category">
    <sch:rule context="f:Observation/f:category/f:coding/f:code">
      <sch:assert test="@value = 'vital-signs'">Vital signs must defined by correct observation code</sch:assert>
      <sch:assert test="count(@value) = 1">Code must exist and be uniquely defined</sch:assert>
    </sch:rule>
    <sch:rule context="f:Observation/f:category/f:coding/f:system">
      <sch:assert test="@value = 'http://terminology.hl7.org/CodeSystem/observation-category'">Vital signs must
        defined by correct system</sch:assert>
    </sch:rule>
  </sch:pattern>
  <sch:pattern id="code">
    <sch:rule context="f:Observation/f:code/f:coding">
      <sch:assert test="f:code/@value = '8867-4'">Heartbeat rate must defined by correct observation code</sch:assert>
      <sch:assert test="f:system/@value = 'http://loinc.org'">Heartbeat rate must defined by correct system</sch:assert>
      <sch:assert test="count(f:code/@value) = 1">Code must exist and be uniquely defined</sch:assert>
    </sch:rule>
  </sch:pattern>
</sch:schema>
```

Figure 50. Schematron for heartbeat rate type data

Line by line explanation of the schematron is as follows:

1. This is the opening tag of the Schematron schema. It declares the start of the schema definition, specifies the XML namespace for Schematron and specifies that the query language used is XSLT 2.0.

2. Defines a namespace with prefix "f" and URI "http://hl7.org/fhir". This namespace will be used in XPath expressions throughout the schema. This namespace is used throughout the schema to specify elements and attributes from the FHIR data format.

3. Defines a namespace with prefix "h" and URI "http://www.w3.org/1999/xhtml". This namespace will also be used in XPath expressions throughout the schema. This namespace is used to specify elements and attributes from the XHTML data format.

4. Line starts a new pattern with ID "observation", which contains rules for validating FHIR Observation resources.

5. Specifies the title of the pattern as "Observation". Provides a human-readable title for the pattern.

6. Starts a new rule within the pattern, which applies to all elements in the document with the FHIR Observation resource type.
7-8. This line defines an assertion (a condition that must be true for the data to be considered valid). It checks that either the dataAbsentReason element is not present, or if it is present, then none of the value elements in the Observation element start with the word "value".
9-11. This line defines another assertion. It checks that either the code element in the component element is not the same as the code element in the Observation element, or if they are the same, then the system attribute in the code element of the component element is not the same as the system attribute in the code element of the Observation element.
12. Defines an assertion that checks that the "id" attribute of the "Observation" element is equal to "heart-rate".
13. Defines an assertion that checks that the "subject" element is present and has a "reference" attribute with a value.
14-15. Defines an assertion that checks that the "effectiveDateTime" element has a value in the correct format.
16. This asserts that an Observation resource must have a measured value, which is represented by the valueQuantity element with a nested value attribute.
17. End of rule definition for Observation resource.
18. End of pattern definition for Observation resource.
19. Declares a pattern with an id of "category". This pattern will contain rules for validating the Observation category code.
20. Declares a rule that applies to the category coding code element within the Observation resource.
21. Declares an assertion that the value of the code element must be "vital-signs".
22. Declares an assertion that the code element must exist and be uniquely defined.
23. End of rule definition for category coding code.
24. Declares a rule that applies to the category coding system element within the Observation resource.
25-26. Declares an assertion that the value of the system element must be http://terminology.hl7.org/CodeSystem/observation-category, i.e., that Vital signs must be defined by the correct system.
27. End of rule definition for category coding system.
28. End of pattern definition for category coding.
29. Declares a pattern with an id of "code". This pattern will contain rules for validating the Observation code.
30. Declares a rule that applies to the code element within the Observation resource.
31. Declares an assertion that the value of the code element must be "8867-4", i.e., Heartbeat rate must be defined by correct observation code.

```
32. Declares an assertion that the system element must have a value of
http://loinc.org, i.e., Heartbeat rate must be defined by the correct system.
33. Declares an assertion that the code element must exist and be uniquely defined.
34. End of rule definition for code.
35. End of pattern definition for code.
36. End of schema.
```

Below is the data after the transformation process, adhering to all the rules in the abovementioned schematron for its particular data type (heartbeat rate).

```
<?xml version="1.0" encoding="UTF-8"?>
<Observation xmlns="http://hl7.org/fhir">
  <id value="heart-rate"/>
  <meta>
    <profile value="http://hl7.org/fhir/StructureDefinition/vitalsigns"/>
  </meta>
  <text> <status value="generated"/> <div xmlns="http://www.w3.org/1999/xhtml">Generated Narrative:
     Observation "heart-rate" </text>
  <category>
    <coding>
      <system value="http://terminology.hl7.org/CodeSystem/observation-category"/>
      <code value="vital-signs"/>
      <display value="Vital Signs"/>
    </coding>
    <text value="Vital Signs"/>
  </category>
  <code>
    <coding>
      <system value="http://loinc.org"/>
      <code value="8867-4"/>
      <display value="Heart rate"/>
    </coding>
    <text value="Heart rate"/>
  </code>
  <subject>
    <reference value="Marija Horvat"/>
  </subject>
  <effectiveDateTime value="2023-03-02"/>
  <valueQuantity>
    <value value="44"/>
    <unit value="beats/minute"/>
    <system value="http://unitsofmeasure.org"/>
    <code value="/min"/>
  </valueQuantity>
</Observation>
```

Figure 51. HL7 compliant heartbeat rate data

Process of validating a XML data against a Schematron results in a report where all, if there are any, rules violations have been found. There are several libraries available for validating XML with Schematron in various programming languages, including Java. One of the most popular Java libraries for Schematron validation is the "Saxon" library, which provides full support for Schematron validation in addition to XSLT and XQuery processing. Saxon is an open-source library that can be used in both commercial and non-commercial applications. To use Saxon for Schematron validation in Java, the following was used:

```
// Load the XML and Schematron files
DocumentBuilderFactory factory = DocumentBuilderFactory.newInstance();
DocumentBuilder builder = factory.newDocumentBuilder();
Document xml = builder.parse(new File("sample.xml"));
Source schematron = new StreamSource(new File("schematron.sch"));

// Create a Schematron validator
Processor processor = new Processor(false);
XsltCompiler compiler = processor.newXsltCompiler();
XsltExecutable exec = compiler.compile(schematron);
XsltTransformer transformer = exec.load();

// Validate the XML
transformer.setSource(new DOMSource(xml));
transformer.setDestination(new NullDestination());
transformer.transform();
```

This code uses the Saxon library to load the XML and Schematron files, create a Schematron validator, and validate the XML against the Schematron rules.

Other programming languages also have libraries available for Schematron validation, such as "libxml2" for C/C++, "lxml" for Python, and "Xerces" for Java and C++. The specific library and code that should be used depend on the programming language and environment of the HIS system (in this case, Java).

If the message contains data as shown on abovementioned Figure 51, the report confirms the data as valid. However, if instead the data (Figure 52) does not comply with all of the rules, the result is shown on Figure 53.

```
<Observation xmlns="http://hl7.org/fhir">
  <id value="heart-rate"/>
  <category>
    <coding>
      <system value="http://terminology.hl7.org/CodeSystem/observation-category"/>
      <code value="vital-signs"/>
    </coding>
  </category>
  <code>
    <coding>
      <system value="http://loinc.org"/>
      <code value="8867-5"/>
    </coding>
  </code>
  <subject></subject>
  <effectiveDateTime value="07-02"/>
  <valueQuantity>
    <value value="44"/>
    <unit value="beats/minute"/>
    <system value="http://unitsofmeasure.org"/>
    <code value="/min"/>
  </valueQuantity>
</Observation>
```

Figure 52. Invalid FHIR heartbeat rate resource

Errors

Severity	Location	Filename	Message
⚠	Line 12	sample.xml	Pattern 'code' Failed : Heartbeat rate must defined by correct observation code.
⚠	Line 15	sample.xml	Pattern 'Observation' Failed : Patient must exist and be uniquely defined.
⚠	Line 16	sample.xml	Pattern 'Observation' Failed : Observation needs to have proper format for dateTime.

Figure 53. Negative report of Schematron validation for heartbeat rate

Upon inspection, it is clear that the observation code does not match the necessary heartbeat rate LOINC code. Furthermore, the patient is not defined. Finally, the datetime field doesn't have a proper format. Thus, the received data is not a proper FHIR resource and does not comply with the specification and HL7 standard.

5.2.2. Body temperature

Conversely, body temperature data is defined by FHIR BodyTemp Structure Definition [154] which is to be adhered to. The structure derives from Observation vital signs. Body temperature data readings are exported in the form of a JSON file. Readings occur every sixty seconds. Each reading consists of a timestamp, the temperature value and the unit used to log the temperature in (in this case, degrees Celsius).

```
<sch:schema xmlns:sch="http://purl.oclc.org/dsdl/schematron" queryBinding="xslt2">
  <sch:ns prefix="f" uri="http://hl7.org/fhir"/>
  <sch:ns prefix="h" uri="http://www.w3.org/1999/xhtml"/>
  <sch:pattern id="observation">
    <sch:title>Observation</sch:title>
    <sch:rule context="f:Observation">
      <sch:assert test="not(exists(f:dataAbsentReason)) or (not(exists(*[starts-with(local-name(.), 'value')])))">dataAbsentReason
        SHALL only be present if Observation.value[x] is not present (inherited)</sch:assert>
      <sch:assert test="not(exists(f:component/f:code)) or
        count(for $coding in f:code/f:coding return parent::*/f:component/f:code/f:coding[f:code/@value=$coding/f:code/@value
        and f:system/@value=$coding/f:system/@value])=0">Component code SHALL not be same as observation code (inherited)</sch:assert>
      <sch:assert test="f:id/@value = 'body-temperature'">Observation is not body temperature type observation</sch:assert>
      <sch:assert test="exists(f:subject/f:reference/@value)">Patient must exist and be uniquely defined</sch:assert>
      <sch:assert test="matches(f:effectiveDateTime/@value, '^(\d{4}-\d{2}-\d{2}T\d{2}:\d{2}:\d{2}\+\d{2}:\d{2})$')">Observation
        needs to have proper format for dateTime</sch:assert>
      <sch:assert test="exists(f:valueQuantity/f:value/@value)">Observation needs to have value measured</sch:assert>
    </sch:rule>
  </sch:pattern>
  <sch:pattern id="category">
    <sch:rule context="f:Observation/f:category/f:coding/f:code">
      <sch:assert test="@value = 'vital-signs'">Vital signs must defined by correct observation code</sch:assert>
      <sch:assert test="count(@value) = 1">Code must exist and be uniquely defined</sch:assert>
    </sch:rule>
    <sch:rule context="f:Observation/f:category/f:coding/f:system">
      <sch:assert test="@value = 'http://terminology.hl7.org/CodeSystem/observation-category'">Vital signs must
        defined by correct system</sch:assert>
    </sch:rule>
  </sch:pattern>
  <sch:pattern id="code">
    <sch:rule context="f:Observation/f:code/f:coding">
      <sch:assert test="f:code/@value = '8310-5'">Body temperature must defined by correct observation code</sch:assert>
      <sch:assert test="f:system/@value = 'http://loinc.org'">Body temperature must defined by correct system</sch:assert>
      <sch:assert test="count(f:code/@value) = 1">Code must exist and be uniquely defined</sch:assert>
    </sch:rule>
  </sch:pattern>
</sch:schema>
```

Figure 54. Schematron for temperature type data

Line by line explanation of the schematron is as follows:

1. This is the opening tag of the Schematron schema. It declares the start of the schema definition, specifies the XML namespace for Schematron and specifies that the query language used is XSLT 2.0.

2. Defines a namespace with prefix "f" and URI "http://hl7.org/fhir". This namespace will be used in XPath expressions throughout the schema. This namespace is used throughout the schema to specify elements and attributes from the FHIR data format.

3. Defines a namespace with prefix "h" and URI "http://www.w3.org/1999/xhtml". This namespace will also be used in XPath expressions throughout the schema. This namespace is used to specify elements and attributes from the XHTML data format.

4. Line starts a new pattern with ID "observation", which contains rules for validating FHIR Observation resources.

5. Specifies the title of the pattern as "Observation". Provides a human-readable title for the pattern.

6. Starts a new rule within the pattern, which applies to all elements in the document with the FHIR Observation resource type.

7-8. This line defines an assertion (a condition that must be true for the data to be considered valid). It checks that either the dataAbsentReason element is not

present, or if it is present, then none of the value elements in the Observation element start with the word "value".
9-11. This line defines another assertion. It checks that either the code element in the component element is not the same as the code element in the Observation element, or if they are the same, then the system attribute in the code element of the component element is not the same as the system attribute in the code element of the Observation element.
12. The first rule in the observation pattern checks that the Observation element has an id attribute with value "body-temperature".
13. The next rule checks that the Observation element has a subject element with a reference attribute that has a unique value. This references the patient.
14-15. The third rule checks that the effectiveDateTime element has a value that matches a specific date-time format.
16. The fourth rule checks that the valueQuantity element has a value attribute with a value.
17. End of rule definition for Observation resource.
18. End of pattern definition for Observation resource.
19. The next pattern with id "category" contains rules for validating the category element within the Observation element.
20. Declares a rule that applies to the category coding code element within the Observation resource.
21. The first rule in the category pattern checks that the code attribute of the coding element within the category element has a value of "vital-signs".
22. The second rule in the category pattern checks that the code attribute of the coding element within the category element is uniquely defined.
23. End of rule definition for category coding code.
24. Declares a rule that applies to the category coding system element within the Observation resource.
25-26. The next rule in the category pattern checks that the system attribute of the coding element within the category element has a specific value.
27. End of rule definition for category coding system.
28. End of pattern definition for category coding.
29. Declares a pattern with an id of "code". This pattern will contain rules for validating the Observation code.
30. Declares a rule that applies to the code element within the Observation resource.
31. Declares an assertion that the value of the code element must be "8310-5", i.e., Body temperature must be defined by correct observation code.
32. Declares an assertion that the system element must have a value of http://loinc.org, i.e., Body temperature must be defined by the correct system.
33. Declares an assertion that the code element must exist and be uniquely defined.

```
34. End of rule definition for code.
35. End of pattern definition for code.
36. End of schema.
```

Below is the data after the transformation process, adhering to all the rules in the abovementioned schematron for its particular data type (body temperature).

```
<?xml version="1.0" encoding="UTF-8"?>
<Observation xmlns="http://hl7.org/fhir">
  <id value="body-temperature"/>
  <meta>
    <profile value="http://hl7.org/fhir/StructureDefinition/vitalsigns"/>
  </meta>
  <text> <status value="generated"/> <div xmlns="http://www.w3.org/1999/xhtml">Generated Narrative:
    Observation "body-temperature"</text>
  <category>
    <coding>
      <system value="http://terminology.hl7.org/CodeSystem/observation-category"/>
      <code value="vital-signs"/>
      <display value="Vital Signs"/>
    </coding>
    <text value="Vital Signs"/>
  </category>
  <code>
    <coding>
      <system value="http://loinc.org"/>
      <code value="8310-5"/>
      <display value="Body temperature"/>
    </coding>
    <text value="Body temperature"/>
  </code>
  <subject>
    <reference value="Marija Horvat"/>
  </subject>
  <effectiveDateTime value="2023-03-02"/>
  <valueQuantity>
    <value value="36.5"/>
    <unit value="C"/>
    <system value="http://unitsofmeasure.org"/>
    <code value="Cel"/>
  </valueQuantity>
```

Figure 55. HL7 compliant body temperature data

In Figure 56, data does not comply with FHIR body temperature resource, as is shown by the rules violation report in Figure 57.

```
<Observation xmlns="http://hl7.org/fhir">
  <id value="body-temperature"/>
  <category>
    <coding>
      <system value="http://terminology.hl7.org/CodeSystem/observation-category"/>
      <code value="not-vital-signs"/>
    </coding>
  </category>
  <code>
    <coding>
      <system value="http://loinc.org"/>
      <code value="8310-5"/>
      <display value="Body temperature"/>
    </coding>
    <text value="Body temperature"/>
  </code>
  <subject>
    <reference value="Patient/example"/>
  </subject>
  <effectiveDateTime value="1999-07-02"/>
  <valueQuantity>
    <unit value="C"/>
    <system value="http://unitsofmeasure.org"/>
    <code value="Cel"/>
  </valueQuantity>
</Observation>
```

Figure 56. Invalid FHIR temperature body resource

Errors

Severity	Location	Filename	Message
⚠	Line 6	sample.xml	Pattern 'category' Failed : Vital signs must defined by correct observation code.
⚠	Line 22	sample.xml	Pattern 'Observation' Failed : Observation needs to have value measured.

Figure 57. Negative report of Schematron validation for body temperature

This time, LOINC code does identify body temperature resources correctly, however, observation code is not set to “vital-signs”. Furthermore, the temperature value itself is missing.

5.2.3. Oxygen saturation

Finally, oxygen saturation in arterial blood is specified by FHIR OxygenSat Structure Definition [155] which is to be adhered to. The structure is developed from Observation vital signs. Oxygen saturation is output in a form of JSON file, consisting of dateTime and value attributes. dateTime has the same format as in previously mentioned heartbeat rate readings while value attribute is expressed in percentage. Figure 58 and Figure 59 present Schematrons for body temperature and oxygen saturation.

```
<sch:schema xmlns:sch="http://purl.oclc.org/dsdl/schematron" queryBinding="xslt2">
  <sch:ns prefix="f" uri="http://hl7.org/fhir"/>
  <sch:ns prefix="h" uri="http://www.w3.org/1999/xhtml"/>
  <sch:pattern id="observation">
    <sch:title>Observation</sch:title>
    <sch:rule context="f:Observation">
      <sch:assert test="not(exists(f:dataAbsentReason)) or (not(exists(*[starts-with(local-name(.), 'value')])))">dataAbsentReason
        SHALL only be present if Observation.value[x] is not present (inherited)</sch:assert>
      <sch:assert test="not(exists(f:component/f:code)) or
        count(for $coding in f:code/f:coding return parent::*/f:component/f:code/f:coding[f:code/@value=$coding/f:code/@value
        and f:system/@value=$coding/f:system/@value])=0">Component code SHALL not be same as observation code (inherited)</sch:assert>
      <sch:assert test="f:id/@value = 'sat02'">Observation is not oxygen saturation type observation</sch:assert>
      <sch:assert test="exists(f:subject/f:reference/@value)">Patient must exist and be uniquely defined</sch:assert>
      <sch:assert test="matches(f:effectiveDateTime/@value, '^(\d{4}-\d{2}-\d{2}T\d{2}:\d{2}:\d{2}\+\d{2}:\d{2})$')">Observation
        needs to have proper format for dateTime</sch:assert>
      <sch:assert test="exists(f:valueQuantity/f:value/@value)">Observation needs to have value measured</sch:assert>
    </sch:rule>
  </sch:pattern>
  <sch:pattern id="category">
    <sch:rule context="f:Observation/f:category/f:coding/f:code">
      <sch:assert test="@value = 'vital-signs'">Vital signs must defined by correct observation code</sch:assert>
      <sch:assert test="count(@value) = 1">Code must exist and be uniquely defined</sch:assert>
    </sch:rule>
    <sch:rule context="f:Observation/f:category/f:coding/f:system">
      <sch:assert test="@value = 'http://terminology.hl7.org/CodeSystem/observation-category'">Vital signs must
        defined by correct system</sch:assert>
    </sch:rule>
  </sch:pattern>
  <sch:pattern id="code">
    <sch:rule context="f:Observation/f:code/f:coding">
      <sch:assert test="f:code/@value = '2708-6'">Oxygen saturation must defined by correct observation code</sch:assert>
      <sch:assert test="f:system/@value = 'http://loinc.org'">Oxygen saturation must defined by correct system</sch:assert>
      <sch:assert test="count(f:code/@value) = 1">Code must exist and be uniquely defined</sch:assert>
    </sch:rule>
  </sch:pattern>
</sch:schema>
```

Figure 58. Schematron for oxygen saturation data

Line by line explanation of the schematron is as follows:

1. This is the opening tag of the Schematron schema. It declares the start of the schema definition, specifies the XML namespace for Schematron and specifies that the query language used is XSLT 2.0.

2. Defines a namespace with prefix "f" and URI "http://hl7.org/fhir". This namespace will be used in XPath expressions throughout the schema. This namespace is used throughout the schema to specify elements and attributes from the FHIR data format.

3. Defines a namespace with prefix "h" and URI "http://www.w3.org/1999/xhtml". This namespace will also be used in XPath expressions throughout the schema. This namespace is used to specify elements and attributes from the XHTML data format.

4. Line starts a new pattern with ID "observation", which contains rules for validating FHIR Observation resources.

5. Specifies the title of the pattern as "Observation". Provides a human-readable title for the pattern.

6. Starts a new rule within the pattern, which applies to all elements in the document with the FHIR Observation resource type.

7-8. This line defines an assertion (a condition that must be true for the data to be considered valid). It checks that either the dataAbsentReason element is not present, or if it is present, then none of the value elements in the Observation element start with the word "value".
9-11. This line defines another assertion. It checks that either the code element in the component element is not the same as the code element in the Observation element, or if they are the same, then the system attribute in the code element of the component element is not the same as the system attribute in the code element of the Observation element.
12. The first rule in the observation pattern checks that the Observation element has an id attribute with value " satO2", identifying Oxygen saturation.
13. The next rule checks that the Observation element has a subject element with a reference attribute that has a unique value. This references the patient.
14-15. The third rule checks that the effectiveDateTime element has a value that matches a specific date-time format.
16. The fourth rule checks that the valueQuantity element has a value attribute with a value.
17. End of rule definition for Observation resource.
18. End of pattern definition for Observation resource.
19. The next pattern with id "category" contains rules for validating the category element within the Observation element.
20. Declares a rule that applies to the category coding code element within the Observation resource.
21. The first rule in the category pattern checks that the code attribute of the coding element within the category element has a value of "vital-signs".
22. The second rule in the category pattern checks that the code attribute of the coding element within the category element is uniquely defined.
23. End of rule definition for category coding code.
24. Declares a rule that applies to the category coding system element within the Observation resource.
25-26. The next rule in the category pattern checks that the system attribute of the coding element within the category element has a specific value.
27. End of rule definition for category coding system.
28. End of pattern definition for category coding.
29. Declares a pattern with an id of "code". This pattern will contain rules for validating the Observation code.
30. Declares a rule that applies to the code element within the Observation resource.
31. Declares an assertion that the value of the code element must be " 2708-6", i.e., Oxygen saturation must be defined by the correct observation code.

32. Declares an assertion that the system element must have a value of http://loinc.org, i.e., Oxygen saturation must be defined by the correct system.
33. Declares an assertion that the code element must exist and be uniquely defined.
34. End of rule definition for code.
35. End of pattern definition for code.
36. End of schema.

Below is the data after the transformation process, adhering to all the rules in the abovementioned schematron for its particular data type (oxygen saturation).

```
<?xml version="1.0" encoding="UTF-8"?>
<Observation xmlns="http://hl7.org/fhir">
  <id value="oxygen-saturation"/>
  <meta>
    <profile value="http://hl7.org/fhir/us/core/StructureDefinition/us-core-pulse-oximetry"/>
  </meta>
  <text> <status value="generated"/><div xmlns="http://www.w3.org/1999/xhtml">Generated Narrative:
     Observation "oxygen-saturation"</text>
  <category>
    <coding>
      <system value="http://terminology.hl7.org/CodeSystem/observation-category"/>
      <code value="vital-signs"/>
      <display value="Vital Signs"/>
    </coding>
    <text value="Vital Signs"/>
  </category>
  <code>
    <coding>
      <system value="http://loinc.org"/>
      <code value="2708-6"/>
      <display value="Oxygen saturation in Arterial blood"/>
    </coding>
    <coding>
      <system value="http://loinc.org"/>
      <code value="59408-5"/>
      <display value="Oxygen saturation in Arterial blood by Pulse oximetry"/>
    </coding>
    <text value="oxygen_saturation"/>
  </code>
  <subject>
    <reference value="Marija Horvat"/>
  </subject>
  <effectiveDateTime value="2023-03-02"/>
  <valueQuantity>
    <value value="99.0"/>
    <unit value="%O2"/>
    <system value="http://unitsofmeasure.org"/>
    <code value="%"/>
  </valueQuantity>
</Observation>
```

Figure 59. HL7 compliant oxygen saturation data

Finally, in Figure 60, non-compliant oxygen saturation XML is given, with the report on Schematron validation given in Figure 61.

```
<Observation xmlns="http://hl7.org/fhir">
  <id value="sat02"/>
  <dataAbsentReason value="data not absent"/>
  <category>
    <coding>
      <system value="http://terminology.hl7.org/CodeSystem/observation-category"/>
      <code value="vital-signs"/>
    </coding>
  </category>
  <code>
    <coding>
      <system value="http://loinc.org"/>
      <code value="2708-6"/>
      <display value="Oxygen saturation in Arterial blood"/>
    </coding>
  </code>
  <subject>
    <reference value="Patient/example"/>
  </subject>
  <effectiveDateTime value="1999-07-02"/>
  <valueQuantity>
    <value value="99.0"/>
    <unit value="%O2"/>
    <system value="http://unitsofmeasure.org"/>
    <code value="%"/>
  </valueQuantity>
</Observation>
```

Figure 60. Invalid FHIR oxygen saturation resource

Errors

Severity	Location	Filename	Message
⚠	Line 2	sample.xml	Pattern 'Observation' Failed : Observation is not oxygen saturation type observation.
⚠	Line 3	sample.xml	Pattern 'Observation' Failed : dataAbsentReason SHALL only be present if Observation.value[x] is not present (inherited).

Figure 61. Negative report of Schematron validation for oxygen saturation

Here, wrong resource ID is given (*sat02* instead of *satO2*). Also, data absent reason is given even though the data is present. Thus, this is not a valid FHIR resource.

The following is a detailed breakdown of the mandatory requirements for the specified data types:

Observation.code must include a single code with:

- fixed value of coding system corresponding to loinc.org
- fixed coding code corresponding to:
    - 8867-4 for heartbeat rate
    - 8310-5 for body temperature
    - 2708-6 for oxygen saturation
- all codes must have system value

Additionally, an Observation must have a value quantity or, in the absence of a value, a reason why the data was not present. In case the value quantity is available, it must have:

- numerical value
- fixed value quantity system corresponding to unitsofmeasure.org
- Unified Code for Units of Measure (UCUM):
    - /min (per minute) for heartbeat rate
    - Cel (Celsius) or degF (Fahrenheit) for body temperature
    - % (per cent) for oxygen saturation

In total, Observation needs to:

- comprise of three mandatory elements (with four more nested mandatory elements),
- support four elements,
- have fixed values for three (body temperature) to four (heartbeat rate, oxygen saturation) elements.

The adoption of the appropriate Schematron ensures adherence to each and every rule for all provided data types. The semantic constraints framework allows the specification of data with typically highly complex structure, while the validation framework can be utilized as an independent component or can be integrated as a module into a major data processing system. General points to note regarding FHIR resource representation:

- FHIR elements are always in the namespace http://hl7.org/fhir, typically defined as default namespace on root element
- Resource names are case-sensitive
- Element names are case-sensitive and must appear in the order the documentation specifies. In case of elements repeating, they must be ordered
- FHIR elements must not be void - they either have a value attribute, valid child element or extension
- Attributes must not be void
- Infrastructural elements need to appear before any other defined child elements, i.e. first base resource elements, then domain resource elements

Element Observation resource offers "measurements and simple assertions made about a patient, device or other subject" and is utilized for vital signs, height, weight, laboratory results, etc. The logical ID of the resource, in this case heartbeat rate, is contained in the element id. Metadata contains information about the resource, and a profile URL, which refers to a structure profile that this resource contends to follow, such as vital signs. An optional Text provides a

human-readable summary of the resource. While coding refers to a code that is explicitly specified by a terminology system (observation-category and http://loinc.org, correspondingly) and distinguished by an existing code, category defines the general sort of observation (vital signs and 8867-4). The display makes the meaning of the code human-readable. The term "subject" refers to the patient, who is the object of the observation. The four types that are accessible for Effective[x] are dateTime, period, timing, and instant. EffectiveDateTime is employed here. DateTime may represent date, date-time, or partial date, with the appropriate formats being YYYY-MM-DD or YYYY-MM-DDThh:mm:ss +zz:zz and YYYY or YYYY-MM, according to the standard. Nonetheless, only date-time is anticipated for this particular use case, and the Schematron rules must reflect this. Value[x] offers information determined as a result of the observation and there are several types possible (as shown in Figure 62). ValueQuantity is utilized here. ValueQuantity is made up of a numerical value, a unit representation (such as bpm), a system that specifies the coded unit form, and the unit itself in coded form. Comparator is a second (optional) argument with the following potential values < (less), less or equal (<=), > (more) and more or equal (>=). In this scenario, mapping data points from the wearable is simple after modifying the datetime format.

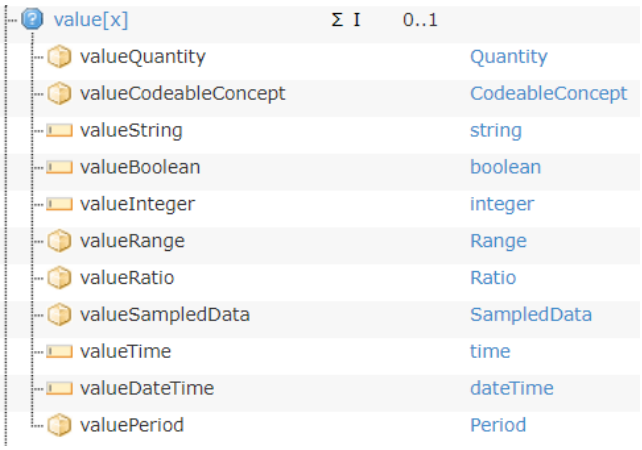

Figure 62. value[x] types

- Since the FHIR Vital Signs profile establishes minimal requirements for the Observation resource to record, search, and get the vital signs, all data types must adhere to the same set of criteria (e.g., temperature, blood pressure, respiration rate, etc.). The regulations relevant to each profile and any extra rules that may be necessary to come after this. Following the guidelines of the most recent FHIR version, all XMLs

generated from data points and consistent with the schematrons established in this research include:data types, in XML and JSON format

- the terminology layer, meaning CodeSystems and ValueSets
- the conformance framework (StructureDefinition)
- the FHIR resources, meaning Patient and Observation

Regarding the implementation described within this research, executable components in a form of parser and validator were developed in Python 3 using Anaconda Spyder, open-source IDE for scientific programming and computing (data science, machine learning applications, large-scale data processing, etc.) and utilizing scientific packages NumPy, SciPy, Matplotlib, pysimdjson and pandas. XML templates and Schematrons were written in Notepad++.

5.3. Integration into existing HIS: conceptual implementation model overview

The choice of integration method often depends on the specific use case and requirements of the healthcare organization. Proposed conceptual model implementation is given in Figure 64 below. Many Hospital Information Systems (HIS) use RESTful APIs (Application Programming Interfaces) to access Electronic Health Record (EHR) data. RESTful APIs are a popular way to allow communication between different software systems over the internet. With RESTful APIs, a HIS can provide access to specific EHR data and functionality to other systems, such as mobile apps, patient portals, and other healthcare applications. RESTful APIs are often used because they provide a flexible and scalable approach to integrating EHR data with other systems. Using RESTful APIs, EHR data can be securely accessed and shared across different systems, allowing healthcare providers to improve patient care, streamline workflows, and enable interoperability between different healthcare systems. Thus, a REST-based access layer for standard HL7-defined data models can be used to enable the FHIR server to receive data generated by wearable sensors. RESTful API is a software architectural style that defines a set of constraints for building web services. RESTful APIs are typically used to expose a set of resources or endpoints that can be accessed over the internet using standard HTTP methods (such as GET, POST, PUT, and DELETE). They are commonly used in modern web applications to provide a scalable and flexible way to access and manipulate data. To use RESTful API with HL7 messages, one approach is to create a RESTful API that can receive and send HL7 messages. This can be done by creating specific endpoints that accept or return HL7 messages in a specific format. For example, the RESTful API could expose an endpoint

that accepts an HL7 message in XML format. Thus, POST endpoint on the server (Figure 63) handles data uploads. Once on the server, parse the incoming request and extract the file and metadata from it. Then, save the file to the server's filesystem or database. Lastly, return a response to the client indicating that the file was successfully uploaded.

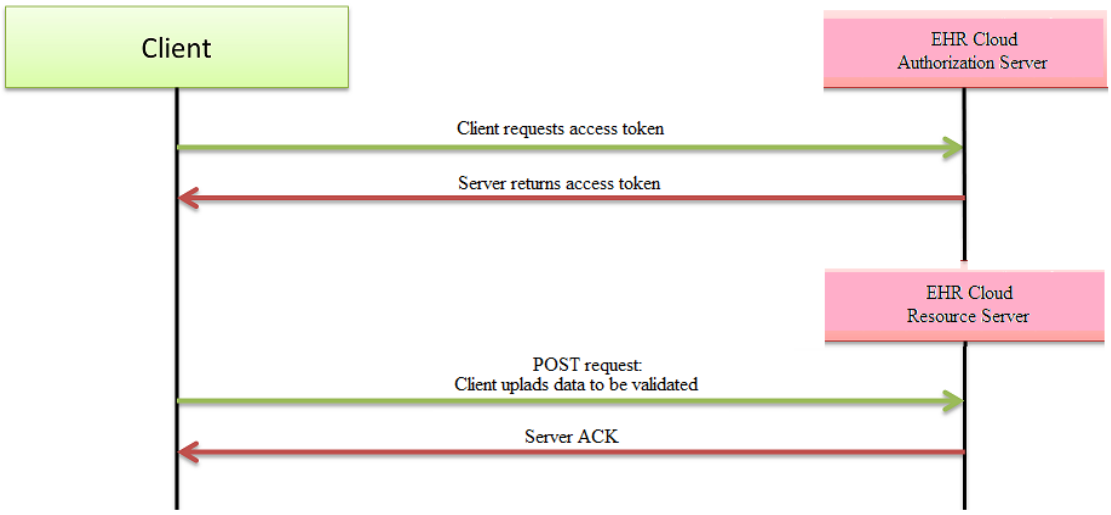

Figure 63. POST REST API

By using RESTful API with HL7 messages, healthcare organizations can create a flexible and scalable solution for exchanging healthcare data. RESTful APIs can provide a modern and user-friendly interface for accessing healthcare data, while HL7 messages can ensure interoperability with legacy healthcare systems. This approach can also provide a more granular level of control over healthcare data, allowing organizations to control which resources are exposed and who has access to them. The HL7 gateway is a middleware that facilitates communication between the HIS and external system granting access to secured FHIR endpoints. HL7 gateways provide a secure and controlled way to exchange healthcare information between different systems. To ensure the confidentiality, integrity, and availability of data transmitted through the HL7 gateway, various authentication and authorization methods can be used, such as Single Sign-On (SSO) authentication.

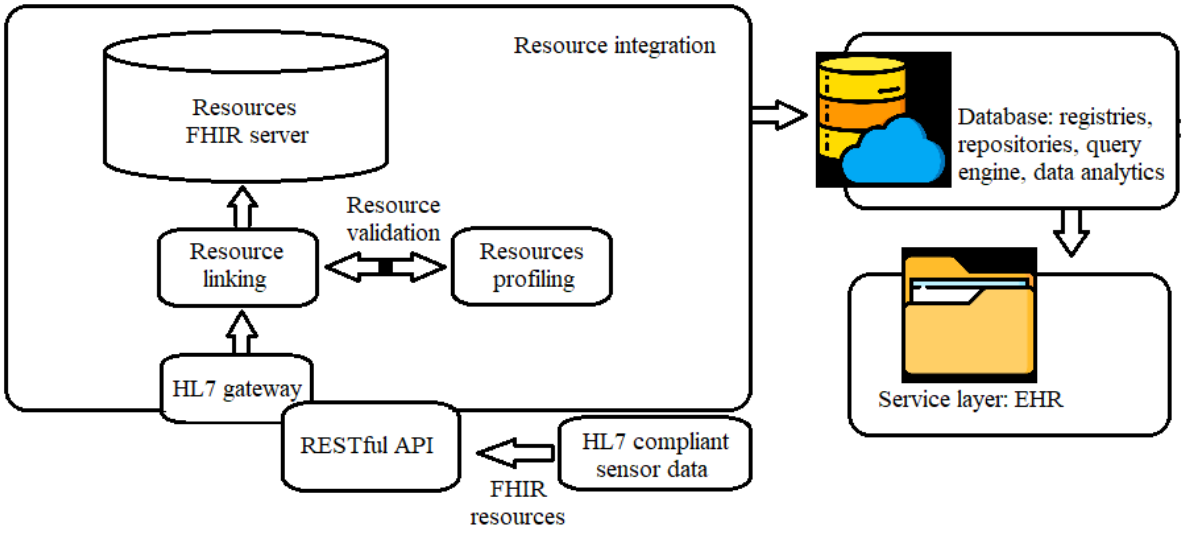

Figure 64. Component for integration of HL7 compliant sensor data into HIS

SSO authentication enables users to authenticate once and then access multiple applications or systems without having to re-enter their credentials. This can be achieved through various authentication protocols such as OAuth 2.0. OAuth 2.0 is widely used in many industries, including healthcare, to provide secure and controlled access to protected resources. OAuth 2.0 is an authorization framework that allows a user to grant a third-party application access to their protected resources on a server without sharing their credentials. In OAuth 2.0, authorization is separated from authentication, which means that a user can grant an application permission to access their data without sharing their login credentials. OAuth 2.0 uses access tokens to grant access to protected resources. These tokens are issued by an authorization server after the user has provided consent. The access token is then sent to the third-party application, which can use it to access the user's protected resources on the server (Figure 65).

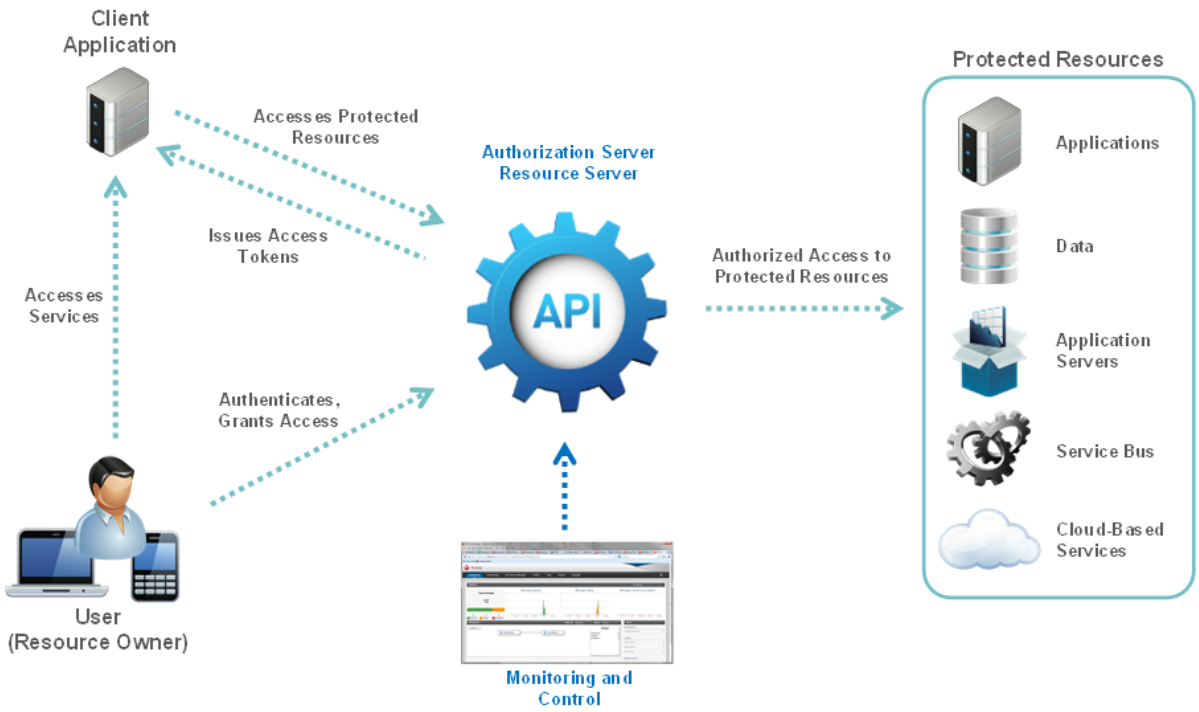

Figure 65. API Gateway as an OAuth 2.0 Resource Server and Authorization Server [199]

Resource profiling involves creating a standardized representation of the sensor data to enable interoperability between different systems. This can be achieved using standard terminologies and data models such as LOINC (Logical Observation Identifiers Names and Codes). By profiling HIS resources using HL7, healthcare organizations can improve the efficiency and effectiveness of data exchange, reduce errors and inconsistencies, and improve patient outcomes by facilitating the sharing of health information across different systems and organizations. The messages are validated with the Schematron for the appropriate data type. Resource linking involves connecting the sensor data with the corresponding patient record in the HIS. This linking can be achieved by using a unique patient identifier or other relevant information such as date of birth or medical record number. FHIR server is a platform that enables the exchange of healthcare data between different systems. The FHIR server stores the standardized sensor data in a FHIR resource format and provides APIs for accessing and exchanging the data. A FHIR server provides a platform for healthcare organizations to store, manage, and share healthcare data in a standardized format that can be easily accessed and integrated with other healthcare applications and systems. FHIR servers can be used to manage various types of healthcare data, including patient information, clinical data, and administrative data. FHIR servers typically support RESTful APIs, which allow healthcare applications and systems to access and exchange healthcare data in a standardized format. FHIR servers also

support various data formats, including XML and JSON, to facilitate interoperability and data exchange across different systems. FHIR servers can be deployed in different ways, including as standalone servers or as part of larger healthcare information systems (HIS). FHIR servers can also be used to support various healthcare use cases, including patient engagement, clinical decision support, and population health management.

6. Security and privacy

Lastly, ensuring privacy and security as the data collected is sensitive in nature; it is crucial to offer a safe system that allows usage of data in an official health information system, — in other words, its integration into EHR. Article [156] has identified two vulnerabilities in communication between activity tracker Fitbit® and its associated Web server:

(a) Cleartext login information pertaining to the user's password, meaning it is sent to the server in clear text and is afterwards logged, and

(b) Cleartext HTTP data processing, meaning the data is transmitted in a form of cleartext as plain HTTP.

Both situations lack data protection and authentication. A comprehensive security study of some of the most widely used fitness tracking devices available is also provided in the article [157]. It focuses primarily on malicious actor behaviors that try to introduce fake data into virtualized cloud services, resulting in inaccurate data analytics. A fraudulent FIT file (reporting 80 million steps taken) was successfully submitted to the cloud using a Man-In-The-Middle (MITM) attack with no warnings being raised. The Advanced Encryption Standard (AES) technique is used for data encryption and decryption in the design for a fitness activity tracking application in [158]. This prevents the manipulation of the data. Encryption and decryption processes are carried out in the database and saved on the server. In summary, the initial task is to reconcile the discrepancies and combine several data sets into a single, coherent aggregate. The effort spent gathering, connecting, and mapping data from various sources turned out to be laborious and time-consuming. Therefore, a transparent, clear, and automated method is required. It would provide useful information and improve the standard of government-based eHealth services (as it would provide a more individualized approach). First, regardless of the source, data must be standardized after being aggregated. This is accomplished by designing schemas. Data schemas define all data types' formats and contents (such as heartbeat rate). This impacts how that data will be processed by software. It is required to transform all data for a single measure (such as heartbeat rate) and integrate it into a standard schema for that distinct metric because the systems in question would be processing data arriving from various devices and platforms, all of which describe the same data differently. A common schema functions as a unique source of documentation as a result, and it may be referred to from any source. The complexity of the data and semantic significance must be taken into account in order to provide reliable insights. Schemas need to specify significant distinctions for every clinical measurement in order to optimize their value and usefulness (e.g., to distinguish whether a heart

rate is at rest or during exercise). The schemas can also be utilized to describe a particular measure (for instance, one value of a blood pressure measurement) or a descriptive statistic of a collection of measurements in general (e.g., an average, minimum or maximum blood pressure value over a period of time). Finally, the usage of schemas standardizes the data and gives it the appropriate clinical context. In addition to the rest of the data from the EHR (such as previous diagnoses and prescriptions, hospitalizations, and laboratory tests), this information can assist doctors in remotely monitoring patients, diagnosing diseases early, and customizing therapies and medications. However, a suitable privacy assurance is essential given the sensitivity of such data. One of the largest challenges to making sure that eHealth solutions are successful is gaining the trust of patients, protecting their privacy, and understanding the ramifications of self-diagnosis without a doctor's assistance. Recent changes aim to provide patients more authority, autonomy and control over decisions involving their personal health information by gradually transferring the control over eHealth data from an institutional level to the level of the patient. Data gathered by using wearable sensors is prone to data security vulnerabilities. With an emphasis on the potential for malicious user insertion of fake data into the tracker's cloud-based services, [157] provides security analysis of a number of fitness trackers that are currently on the market. As data is wirelessly sent from the sensor to the smartphone via Bluetooth LE (low energy), [159] highlights the significance of secure pairing protocols to prevent eavesdropping attempts. By executing attacks to find weaknesses in hardware and software, SecuWear [160], a multiple-domain wearables testbed framework, improves the security investigation of wearables. The greatest threat to data privacy in cloud-based applications is recognized as unencrypted communication [161]. SensCrypt, a secure protocol for Bluetooth activity tracker management, is presented by [162] and [163] suggests a filtering system that balances the security and sharing of health data. The AES algorithm is used by [158] for data encryption and decryption because it prevents data manipulation. Another concern is privacy, since various health-related applications that connect with well-known activity trackers interact with "unexpected" third parties such as social networks [18]. Even though the application's privacy policy is required to make this information clear, the majority of users never review it [164], therefore they are unaware that their data is being shared. Finally, for better, more individualized, and more economical healthcare, it is essential that people accept modern healthcare technology. Evidence shows that some patients, however, object because they feel their privacy is being invaded. [165] focuses on how privacy issues affect the choice of whether to use eHealth technology. The findings indicate that the perception of advantages is the best predictor of the usage of digital health technology. Similar to this, [166]

demonstrates a strong association between people's ongoing health conditions and their decisions to adopt healthcare technology.

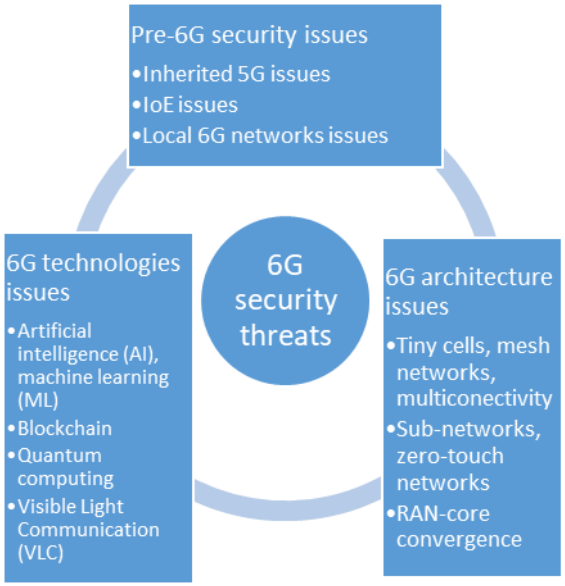

Figure 66. Security and privacy threats in the context of 6G

Figure 66 illustrates an analysis on 6G's security and privacy issues. It covers:

- Attacks on weaker cryptographic systems by the use of quantum computers,
- Corrupted or hijacked IoT devices,
- Data theft from IoT devices,
- Eavesdropping,
- Signal jamming,
- Node compromise attack on Visible Light Communication (VLC).

Artificial intelligence is considered as the major technology to help reduce these dangers [167]. Additionally, [168] suggests a BloCoV6, a blockchain-aided unmanned aerial vehicles (UAV) contact-tracing strategy, for locating prospective COVID-19 patients. Despite the difficulties, there is massive interest in this system's potential and advantages, particularly in light of the present coronavirus (COVID-19). Possible use-case scenarios for telemedicine and wearable technology during the COVID-19 pandemic are shown in Figure 67.

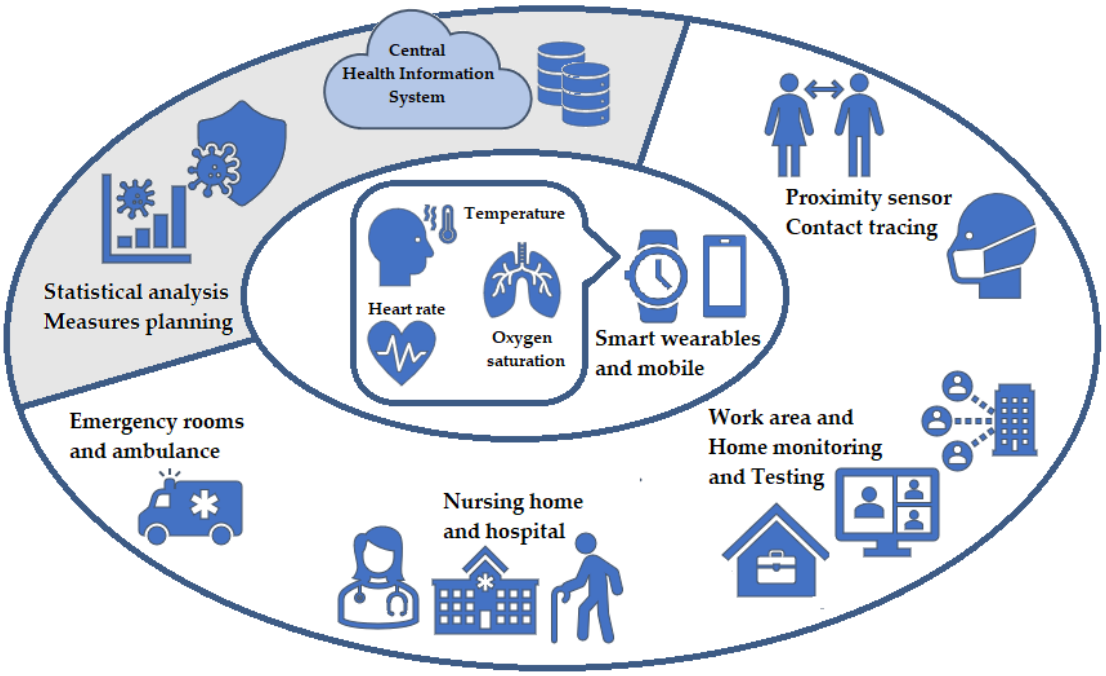

Figure 67. Use-case scenarios for sensor-equipped wearables and telehealth

6.1. Security and privacy threats

Data security safeguards digital data from being accessed by unauthorized users, altered in a malicious manner (corrupted), or stolen. The following are three significant dangers to the security of medical data [161]:

- integrity - data must not be altered (tampering),
- availability - it must be possible to readily access data when needed,
- confidentiality - data must not be shared with an unauthorized party.

Privacy guarantees that data is handled correctly, including obtaining users' consent, providing relevant notice, and adhering to legislative requirements. Particular worries in relation to data privacy include:

- sharing data with potentially unauthorized third party,
- data collection and storage methods,
- regulatory constraints (e.g., GDPR or HIPAA).

Regulation (GDPR), a single piece of legislation. Health Insurance Portability and Accountability Act is the legislation that governs medical data protection in the United States (HIPAA). In Table 9 below, a comparison between GDPR and HIPAA is provided.

Table 9. GDPR and HIPAA comparison

	GDPR	HIPAA
Data protected	Any non-anonymized data related to identifiable individual	Protected Health Information (PHI)
Accountability	Data controller	Covered entity (healthcare provider)
Breach notification	The breach has to be reported within 72 hours; the users affected must be informed	Health provider must notify the patient
Third parties	Written safeguards	User must be informed
Sanctions	Depends on the country	Criminal and money penalties

In this research, the focus is on the EU and GDPR. General principles of GDPR are:

- consent – patient needs to be fully informed and must express their explicit consent to processing of data,
- purpose limitation – purposes must be clear and relevant (impactful),
- data minimization – gathering, using and storing only necessary data,
- transparency to the patient,
- accuracy of the gathered data,
- privacy by design (default) – taking into account privacy during the planning and implementation processes,
- data subject rights – patients have the right to seek data deletion and to view their data; data retention duration - information shouldn't be kept forever, data integrity, availability, and confidentiality are ensured by security measures. "Personal data relating to physical or mental health of a person, including the provision of health care services, which reveal information about their health state," is how GDPR's Article 4 paragraph 15 defines "health data." Some sorts of data, such those from activity trackers, aren't always included in this category. Health information is therefore further defined as: refers to information that is obtained in a professional medical environment, such as EHR dataRaw data, such as that gathered by fitness tracker sensors, only in cases it is used to evaluate a person's health. Step count may or may not be considered health data depending on whether it is being utilized in a medical setting or not, in contrast to data like cardiac rate, oxygen saturation, or blood pressure, which are clearly identified as such.

Because different categories of data may have distinct legal ramifications, it is necessary to clearly specify them when using raw data from a tracker.

When managing private data, a Privacy Impact Assessment (PIA) is an essential step. The purpose of PIA is to identify and evaluate privacy issues while gathering, storing, managing, and exchanging data in order to reduce the risks associated with information [169]. Before beginning the creation of any system that is intended to gather or handle the personal data of persons, an evaluation must be done. So that the risks may be further examined, any privacy concerns discovered throughout this procedure must be reported. Then, mandated activities are established in accordance with the impact and likelihood that have been identified. Figure 68 provides a summary of the security and privacy risks that have been found in the eHealth system that leverages patient-side obtained personal health data, such as information from a wearable device; it shows how handling personal health data can be subject to security and privacy risks at the subsequent steps:

- data collection (sensors) and communication to mobile devices;
- data transmission via wireless networks;
- processing and storing information on healthcare servers, i.e., adhering to standards and regulations; and
- accessing stored information (e.g., physician viewing patient's EHR).

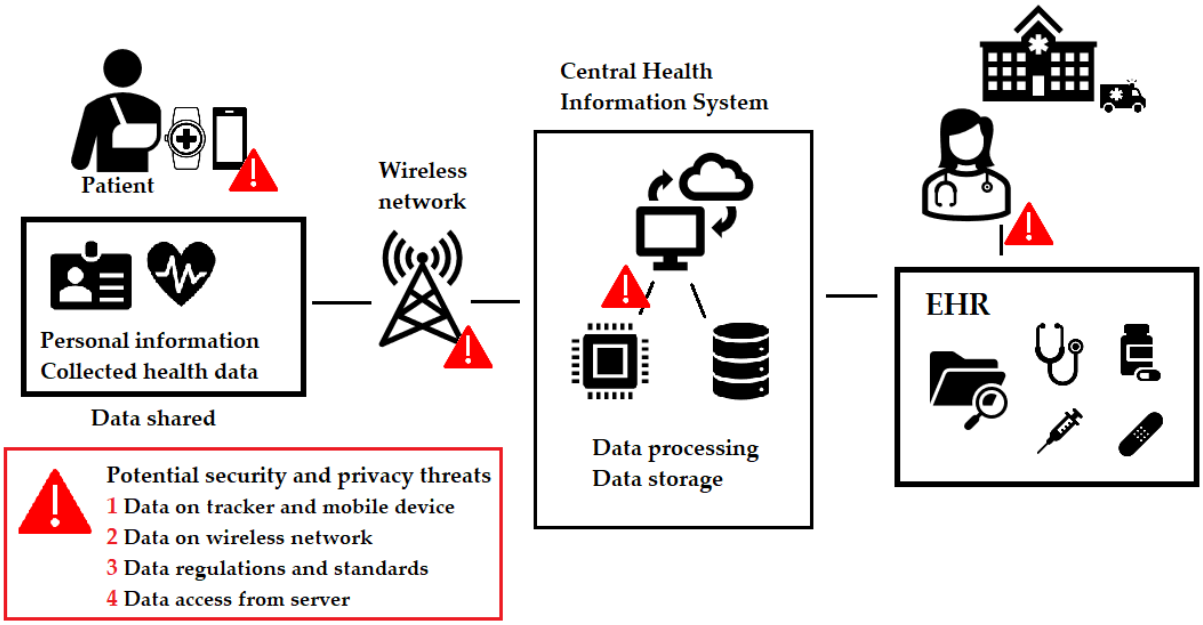

Figure 68. Overview of security and privacy issues in mHealth system

6.1.1. Data on tracker and mobile devices

Fitness trackers have been found to have a number of vulnerabilities over the years, including the following:

- using third party analysis tools,

- lack of privacy policies,
- internal (device) or external (cloud) insufficient encryption,
- insufficient security at transport layer, for example, using HTTP instead of HTTPS protocol,
- insufficient security at application level, such as allowing injection on the client side, e.g., SQL injection,
- poor authentication and authorization (password-protected access),
- incorrect session handling.

These may be avoided, but while building activity trackers, producers must assure privacy and security. Any mHealth mobile application must also be impervious to manipulation with data and able to defend itself by identifying threats in real time.

6.1.2. Data on wireless network

Data transfer between a mobile device and a cloud server is vulnerable to manipulation, including Man-in-the-Middle attacks (MITM). End-to-end encryption (E2EE), shown in Figure 69, using a device-specific key is required to avoid this from happening.

6.1.3. Data regulations and standards

Data must be kept in a manner that complies with all applicable standards and regulations, including healthcare data type semantic constraints. Verifying and validating data utilizing a Schematron-based validation process can do this [170]. Legal requirements must also be followed. The GDPR requires privacy policies that expressly warn patients about the data that will be collected, how it will be used, and whether it will be shared with outside parties. Data usage requires patient consent, which must be obtained. The patient has the right to rescind their permission at any moment.

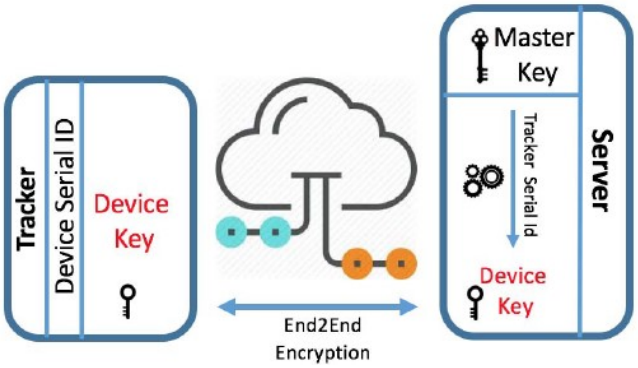

Figure 69. End-to-end encryption

6.1.4. Data access from server

As it is used in several Central HIS throughout the world, the EHR demonstrates its value as a tool in the delivery of medical treatment. Security procedures must be put in place within the medical facility to prevent unauthorized individuals from entering it (or other access point). In order to protect patient privacy and secure the security of data inside the EHR, the policy must contain role-based access control. Table 10. concludes by outlining potential security and privacy threats as well as mitigating measures.

Table 10. List of possible risks and solutions

Risk	Solution	Result
Purpose of data collecting has not been explained to the patient	The purpose must be explicitly stated in the Privacy policy or in a consent form signed by the patient	Mitigated
The nature of data is not explicitly stated to the patient (i.e., exact data types)	Data types must be explicitly stated in the Privacy policy or in a consent form signed by the patient (e.g., heartbeat, oxygen saturation)	Mitigated
Extent of data usage not explicitly stated	Extent to which data will be used must be explicitly stated in Privacy policy or consent form signed by the patient	Mitigated
Potential sharing of data with third parties not addressed	In case of sharing the data with third parties, must state this beforehand in the Privacy policy or consent form signed by the patient	Mitigated
Patient has no option to withdraw consent at any time and have their data deleted	Patient must be able to revoke their consent at any time at which point all their data must be deleted	Mitigated
Patient not able to revise or correct data they deem incorrect	Secure channel must exist through which patients can inform errors in data exist and data revision or modification is necessary	Mitigated
Data is not anonymized	Data must be anonymized, i.e., processed in such a way	Mitigated

	that makes identification of a specific patient impossible	
Data stored longer than it is necessary for processing	Data is kept only temporarily during processing until it is uploaded to the cloud	Mitigated
Data is not stored securely (e.g., data is publicly accessible or unencrypted)	Data must be stored encrypted and mustn't be accessible from external storage devices or applications	Mitigated
Password or encryption keys are kept in plain text	Secure Hash Algorithm (SHA) is used to store keys and passwords	Mitigated
Insecure points of input; risks of client-side attacks or data tampering (e.g., SQL injection)	Prevented by input sanitization	Mitigated
Insecure data transmission channel	Data transmission secured by Secure Sockets Layer (SSL) protocol	Mitigated
Inadequate logging practices, i.e., writing sensitive data into logs	Ensure no confidential data is written into logs	Mitigated
Data backup lacks encryption	Data backup stored after encryption	Mitigated
Data breaches, i.e., access by unauthorized parties	Authentication and authorization process in place, following AuthO2 standard. Role-based access control	Mitigated
Data tampering, i.e., modifications to data made by an unauthorized party	All data is encrypted, all keys are securely hidden	Mitigated

Possibility of exploits by malicious software, i.e., taking advantage of bugs and vulnerabilities to cause harm	Sandboxed software, i.e., isolated virtual machine	Partially mitigated

7. Discussion

Wireless sensor networks commonly experience faults, which can accumulate and decrease the accuracy of sensory readings over time. This can negatively impact the reliability of the network and lead to incorrect conclusions. To use personal health data in a machine learning algorithm, data quality assessment and cleaning are necessary. This involves assessing specific data quality indicators such as accuracy, timeliness, completeness, and consistency, and using similarity and correlations between attributes to identify incomplete or corrupt data. In Chapter 4, data-driven models for data cleaning were compared on a use case of ECG signal, namely, multiple linear regression, decision tree, random forest, artificial neural network and support vector machines. In one of the leads' data, 10% of the data was discarded, and the missing values were computed using statistical and machine learning techniques in order to compare the quality of computation of various data-driven models. The estimated values are then contrasted with the actual values, and models are assessed by contrasting the calculated errors. Multiple linear regression and neural networks have shown the best results and were chosen for further optimization. The missing data were imputed using the correlation and similarity. This method can be used on a wide range of data, and as a result, it has numerous potential uses. RMSE and RRMSE were calculated to quantify the difference between the two datasets. This step provides a measure of how much the cleaning process changed the original dataset and how well the imputation algorithm performed in filling in the missing data. By randomly marking some data as erroneous and then using a machine learning algorithm to impute missing values, the cleaned dataset can be compared to the original dataset to evaluate the effectiveness of the cleaning process. The goal was to optimize the data cleaning model and improve accuracy by using an extra variable - intensity of physical activity, which is known to correlate with the values being imputed. These imputed entries were then combined into a complete dataset that has no missing information. For combined datasets, the accuracy increased by a total of 10% to 17%. This has resulted in a system model for data cleaning and transformation based on a combination of selected classification and regression models and their optimization which ensures accuracy of data collected.

The process described in Chapter 5 was used to model a validation process that made sure the data adhered to regulations and standards. This was achieved by:

- Defining semantic constraints for healthcare data types in order to guarantee adherence to standards and regulations making the information valid and relevant from a medical standpoint

- Defining and modeling a procedure for validating the obtained data so that it may be readily transferred and included into an official EHR. Lastly, this methodology was tested in a use-case study utilizing an existing dataset that contained a variety of pertinent data types.
- Additionally, conceptual implementation model overview for integration into an existing HIS was proposed.

For the purpose of using personal health data in individualized and preventative healthcare, compliance with regulations and standards is crucial. By using the procedure outlined in this work, newly processed data may be officially added to an EHR in accordance with IHE standards and regulations for the relevant data types. Finally, Table 8 below provides a comparison of the suggested model with those already described in literature. The model is consistent with the leading industry standard, has its focus on integrating data into EHR, offers a data cleaning module and automatic Schematron-based validation.

Table 8. Comparison of the performances and functionalities of various system models

	Data collected via wearables	Data cleaning process	Data integration	Validation process	Compliance with standards
WearableHUB	Yes	N/A	PHR	N/A	N/A
Angel-Echo	Yes	N/A	PHR	N/A	N/A
OpenHealth	Yes	Machine learning algorithms	PHR	N/A	N/A
mHealth	No	N/A	PHR/EHR	N/A	HL7 FHIR-based
Tangle	Yes	N/A	PHR/EHR	N/A	HL7 FHIR-based
mHealth4Afrika	Yes	N/A	PHR/EHR	Validation through human review; group interviews and observation	HL7 FHIR-based
Proposed model	Yes	Machine learning algorithm: neural network	EHR focus	Schematron validation	HL7 FHIR-based

In order to guarantee compliance with standards and regulations and to make the information medically useful and valid, semantic constraints for healthcare data types were developed. It is suggested to use a semantic validation and Schematron-based validation procedure. The data

will be able to be transferred and included into a formal EHR due to the validation method provided in this research. The method was subsequently validated using data sets that include several health-related data types.

8. Conclusions

Given that the accuracy of sensory measurements declines over time, faults in wireless sensor networks are very prevalent. Therefore, if not adequately addressed, the buildup of flaws can have a major detrimental impact on the network's dependability and result in false inferences. To be used in a machine learning algorithm, the quality of the collected personal health data (a variety of person's metrics and vital signs, including heartbeat rate, oxygen saturation, respiration rate, body temperature, electrocardiogram, electromyogram, blood pressure, blood glucose levels, and galvanic skin response) must be sufficient. In this situation, the required process entails the estimation of data quality and data cleaning process, which entails breaking down the data quality into distinct data quality indicators, these being accuracy or correctness, timeliness, completeness, and consistency. This is done via use of similarity and correlations between attributes, i.e., by creating co-appearance matrices, normalized similarity matrix for attributes, and correlation matrix. The correlation between physiological characteristics is well established, and alterations in at least two or more indicators are common (e.g., the heart rate and respiration ratio increase simultaneously). A precise missing value imputation (e.g., statistical pattern recognition, especially, decision trees and regression) is essential to boost the usability of the data set for continued usage in a HIS when incomplete or corrupt data has been found.

Various data-driven models for imputing data were examined as a part of this research for a diverse range of subjects, and the findings indicated that neural networks and multiple linear regression were the most effective. These algorithms were picked for improvement as a result. Additionally, the data was afterwards categorized according to the level of activity performed and imputed piece by piece, producing better imputation results.

Thus, the effort contributes the following:

- Thorough comparison of data-driven methodologies to identify the best imputation model for future advancements in calculating ECG signal data obtained from sensors.
- An improved method in which physical activity is first separated from the health-related sensor data before being categorized and then imputed. The suggested data-driven approach has demonstrated 10-17% better predictions by integrating classification and regression methods.

Furthermore, by deliberately attempting to check incorrect XML files against the relevant Schematron, conformance to every rule for all provided data types has been thoroughly

validated. The research's next major obstacle was defining a validation procedure to make sure the data complied with standards. This was accomplished by:

- Defining semantic restrictions for health-related data types to assure standard compliance and make the information meaningful and legitimate from a medical standpoint
- Defining and modeling the validation process for the gathered data, allowing for simple transfer and incorporation of the data into a formal EHR. Last but not least, this methodology was tested in a use-case study utilizing an existing dataset that contained a variety of pertinent data types.

To utilize personal health information for tailored and preventative treatment, compliance with standards is necessary. By using the procedure outlined in this work, produced data can be added to a formal EHR while adhering to the most recent IHE standards for the specified data types. The model is consistent with the top industry standard, focuses on EHR data integration, offers a data cleaning module, and offers automatic Schematron-based validation.

The work also provides in-depth identification and analysis of security and privacy threats when integrating IoMT health data into EHR in both the pre-6G and future 6G eras. Pre-6G risks are addressed with solutions, while 6G general solution concepts are examined, with certain outstanding questions being noted. This completes the concept of a system architecture for the integration of data from wearable smart devices inside the central health information system, including harmonization and validation of data with global standards and regulations pertaining to the EHR.

References

[1] J. Andreu-Perez, D. R. Leff, H. M. D. Ip and G. -Z. Yang, "From Wearable Sensors to Smart Implants-–Toward Pervasive and Personalized Healthcare," in IEEE Transactions on Biomedical Engineering, vol. 62, no. 12, pp. 2750-2762, Dec. 2015, doi: 10.1109/TBME.2015.2422751.

[2] M. Viceconti, P. Hunter and R. Hose, "Big Data, Big Knowledge: Big Data for Personalized Healthcare," in IEEE Journal of Biomedical and Health Informatics, vol. 19, no. 4, pp. 1209-1215, July 2015, doi: 10.1109/JBHI.2015.2406883.

[3] Global Industry Analysts, Inc. (June 2020). Fitness Bands - Global Market Trajectory & Analytics.

[4] Mordor Intelligence. (January 2020). Smart Watch Market - Growth, Trends, and Forecast (2020 - 2025).

[5] Grand View Research, Inc. (2019). Internet of Things in Healthcare Market Size, Share & Trends Analysis Report by Component, by Connectivity Technology, by End Use, by Application, And Segment Forecasts, 2019 - 2025.

[6] Markets and markets. (2020). IoT in Healthcare Market by Component (Medical Device, Systems & Software, Services, and Connectivity Technology), Application (Telemedicine, Connected Imaging, and Inpatient Monitoring), End User, and Region - Global Forecast to 2025.

[7] World Health Organization. (2021, December 20th). WHO. Retrieved from Covid-19: https://covid19.who.int/

[8] Chamola, V., Hassija, V., Gupta, V., & Guizani, M. (2020). A Comprehensive Review of the COVID-19 Pandemic and the Role of IoT, Drones, AI, Blockchain, and 5G in Managing its Impact. IEEE Access, vol. 8, pp. 90225-90265. doi:10.1109/ACCESS.2020.2992341.

[9] N. Ahmed, R. A. Michelin, W. Xue, S. Ruj, R. Malaney, S. S. Kanhere, A. Seneviratne, W. Hu, H. Janicke, S. K. Jha. (2020). A Survey of COVID-19 Contact Tracing Apps. IEEE Access, vol. 8, pp. 134577-134601. doi:10.1109/ACCESS.2020.3010226

[10] Sahraoui, Y., Kerrache, C. A., Korichi, A., Nour, B., Adnane, A., & Hussain, R. (2020). DeepDist: A Deep-Learning-Based IoV Framework for Real-Time Objects and Distance Violation Detection. IEEE Internet of Things Magazine, vol. 3, pp. 30-34. doi:10.1109/IOTM.0001.2000116

[11] Kim, N., Wei, J. L., Ying, J., Zhang, H., Moon, S. K., & Choi, J. (2020). A Customized Smart Medical Mask for Healthcare Personnel. 2020 IEEE International Conference on

Industrial Engineering and Engineering Management (IEEM), (pp. pp. 581-585). doi:10.1109/IEEM45057.2020.9309863

[12] Kalavakonda, R. R., Masna, N. V., Bhuniaroy, A., Mandal, S., & Bhunia, S. (2021, March 1st). A Smart Mask for Active Defense Against Coronaviruses and Other Airborne Pathogens. IEEE Consumer Electronics Magazine, vol. 10, pp. pp. 72-79. doi:10.1109/MCE.2020.3033270

[13] Vishnu, S., & Ramson, S. R. (2021). An Internet of Things Paradigm: Pandemic Management (incl. COVID-19). 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), (str. pp. 1371-1375). doi:10.1109/ICAIS50930.2021.9395966

[14] Nadeem, O., Saeed, M. S., Tahir, M. A., & Mumtaz, R. (2020). A Survey of Artificial Intelligence and Internet of Things (IoT) based approaches against Covid-19. 2020 IEEE 17th International Conference on Smart Communities: Improving Quality of Life Using ICT, IoT and AI (HONET), (str. pp. 214-218). doi:10.1109/HONET50430.2020.9322829

[15] Colaco, J., & Lohani, R. B. (2020). Health Care System for Home Quarantine People. 2020 IEEE 1st International Conference for Convergence in Engineering (ICCE), (str. pp. 147-151). doi:10.1109/ICCE50343.2020.9290557

[16] Panicacci, S., Giuffrida, G., Donati, M., Lubrano, A., Ruiu, A., & Fanucci, L. (2021). Empowering Home Health Monitoring of Covid-19 Patients with Smartwatch Position and Fitness Tracking. 2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS), (pp. pp. 348-353). doi:10.1109/CBMS52027.2021.00109

[17] Zhang, T., Liu, M., Yuan, T., & Al-Nabhan, N. (2021). Emotion-Aware and Intelligent Internet of Medical Things towards Emotion Recognition during COVID-19 Pandemic. IEEE Internet of Things Journal. doi:10.1109/JIOT.2020.3038631

[18] Kazlouski, A., Marchioro, T., Manifavas, H., & Markatos, E. (2021). Do partner apps offer the same level of privacy protection? The case of wearable applications. 2021 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops), (str. pp. 648-653). doi:10.1109/PerComWorkshops51409.2021.9431018

[19] Xiaoxia, Q. (2020). How emerging technologies helped tackle COVID-19 in China. World Economic Forum 2020.

[20] Hu, J., Zhang, H., Di, B., Li, L., Bian, K., Song, L., Poor, H. V. (2020). Reconfigurable Intelligent Surface Based RF Sensing: Design, Optimization, and Implementation. IEEE Journal on Selected Areas in Communications, vol. 38, pp. 2700–2716.

[21] Nguyen, D., Ding, M., Pathirana, P. N., & Seneviratne, A. (2020). Blockchain and AI-based Solutions to Combat Coronavirus (COVID-19)-like Epidemics: A Survey.

[22] Mucchi, L., Jayousi, S., Caputo, S., Paoletti, E., Zoppi, P., Geli, S., & Dioniso, P. (2020). How 6G Technology Can Change the Future Wireless Healthcare. Proceedings of the 2020 2nd 6G Wireless Summit (6G SUMMIT), (pp. pp. 1–6.).

[23] ABI Research. (2021). 6G Standards and Market Developments: Application Analysis Report.

[24] Sinha, P., Sunder, G., Bendale, P., Mantri, M., & Dande, A. (2013). Electronic Health Record: Standards, Coding Systems, Frameworks, and Infrastructures. IEEE Press.

[25] Health informatics — Capacity-based eHealth architecture roadmap — Part 2: Architectural components and maturity model, ISO/TR 14639-2:2014, Geneva, Switzerland: ISO.

[26] P. Szolovits, J. Doyle, W. J. Long, I. Kohane, S. G. Pauker, “Computerisation of personal health records,” Health Visitor, vol. 51, no. 6, p. 227, Jun. 1978.

[27] P. Szolovits J. Doyle, W. J. Long, I. Kohane, S. G. Pauker, “Guardian Angel: Patient-centered health information systems,” MIT Lab. Comput. Sci., Cambridge, MA, USA, TR-604, May 1994.

[28] W. W. Simons, K. D. Mandl, and I. S. Kohane, “The PING personally controlled electronic medical record system: Technical architecture,” J. Amer. Med. Inform. Assoc., vol. 12, no. 1, pp. 47–54, Jan. 2005.

[29] A. Brown and B. Weihl, “An update on Google Health and Google PowerMeter,” Jun. 2011. https://googleblog.blogspot.com/2011/06/update-on-google-health-and-google.html, accessed November 2022.

[30] M. L. Braunstein, "Health care in the age of interoperability part 5: the personal health record," in IEEE Pulse, vol. 10, no. 3, pp. 19-23, May-June 2019, doi: 10.1109/MPULS.2019.2911804.

[31] Agencies (2018-03-30). "Hackers steal data of 150 million MyFitnessPal app users", The Guardian, https://www.theguardian.com/technology/2018/mar/30/hackers-steal-data-150m-myfitnesspal-app-users-under-armour, accessed November 2022

[32] Weed LL. Medical records that guide and teach. N Engl J Med. 1968 Mar14;278(11):593–600.

[33] McDonald CJ, Tierney WM. The Medical Gopher – a microcomputer system to help find, organize and decide about patient data. West J Med 1986; 145 (6): 823–9.

[34] Evans, R. S. (2016). Electronic Health Records: Then, Now, and in the Future. Intermountain Healthcare & Biomedical Informatics, University of Utah School of Medicine, Salt Lake City, USA.

[35] Salenius, S., Margolese-Malin, L., Tepper, J., Rosenman, J., Varia, M., & L., H. (1992). An electronic medical record system with direct data-entry and research capabilities. International Journal of Radiation Oncology - Biology - Physics (IJROBP).

[36] Harrington, J. (1991). Application of open systems to health care communications. IEEE Medical Data Interchange (MEDIX).

[37] Wen, H., Ho, Y., Jian, W., Li, H., & Hsu, Y. (2007). Scientific production of electronic health record research, 1991-2005. Comput Methods Prograns, BioMed.

[38] L. L. Fragidis and P. D. Chatzoglou, "Implementation of a nationwide electronic health record (ehr) the international experience in 13 countries," International journal of health care quality assurance, vol. 31, no. 2, pp. 116–130, 2018.

[39] J. C. Mandel, D. A. Kreda, K. D. Mandl, I. S. Kohane, and R. B. Ramoni, "Smart on FHIR: a standards-based, interoperable apps platform for electronic health records," Journal of the American Medical Informatics Association, vol. 23, no. 5, pp. 899–908, 2016.

[40] K. B. Eden, A. M. Totten, S. Z. Kassakian, P. N. Gorman, M. S. McDonagh, B. Devine, M. Pappas, M. Daeges, S. Woods, and W. R. Hersh, "Barriers and facilitators to exchanging health information: a systematic review," International journal of medical informatics, vol. 88, pp. 44–51, 2016.

[41] Health Level 7 Fast Healthcare Interoperability Resources. https://www.hl7.org/fhir/, accessed September 2022

[42] Fennell, P. (2014). Schematron - More useful than you'd thought. XML London 2014, (str. pp. 103–112).

[43] Integrating Health Enterprise, IHE. https://www.ihe.net/about_ihe/, accessed September 2021

[44] IHE Resources, Technical Frameworks. https://www.ihe.net/resources/technical_frameworks/#IT, accessed December 2022

[45] IHE ITI Technical Framework Volume 1 Integration Profiles. https://www.ihe.net/uploadedFiles/Documents/ITI/IHE_ITI_TF_Vol1.pdf, accessed December 2022

[46] IHE ITI Technical Framework Volume 2a Transactions, 2021

[47] IHE ITI Technical Framework Volume 2b Transactions. https://www.ihe.net/uploadedFiles/Documents/ITI/IHE_ITI_TF_Vol2b.pdf, accessed December 2022

[48] IHE ITI Technical Framework Volume 2a Transactions. https://www.ihe.net/uploadedFiles/Documents/ITI/IHE_ITI_TF_Vol2a.pdf, accessed December 2022

[49] IHE ITI Technical Framework Volume 3 Metadata. https://www.ihe.net/uploadedFiles/Documents/ITI/IHE_ITI_TF_Vol3.pdf, accessed December 2022

[50] IHE ITI Metadata Update. https://www.ihe.net/uploadedFiles/Documents/ITI/IHE_ITI_Suppl_XDS_Metadata_Update.pdf, accessed December 2022

[51] IHE ITI Cardiology Technical Framework Resting ECG Workflow. https://www.ihe.net/uploadedFiles/Documents/Cardiology/IHE_CARD_Suppl_REWF.pdf, accessed December 2022

[52] IHE ITI Cardiology Technical Framework Stress Testing Workflow. https://www.ihe.net/uploadedFiles/Documents/Cardiology/IHE_CARD_Suppl_Stress.pdf, accessed December 2022

[53] IHE ITI Quality, Research and Public Health Technical Framework Supplement Aggregate Data Exchange. https://www.ihe.net/uploadedFiles/Documents/QRPH/IHE_QRPH_Suppl_ADX.pdf, accessed December 2022

[54] Specifications Editorial Committee. "openEHR EHR Extract IM". openEHR Foundation.

[55] openEHR Specification Program. "openEHR Specifications". openEHR Foundation.

[56] openEHR Specification Program. "openEHR Architecture Overview". openEHR Foundation. openEHR Foundation.

[57] openEHR Specification Program. "Archetype Technology Overview". openEHR Foundation.

[58] openEHR Specification Program. "Archetype Query Language (AQL)". openEHR Foundation.

[59] G. M. Bacelar-Silva, H. César, P. Braga and R. Guimarães, "OpenEHR-based pervasive health information system for primary care: First Brazilian experience for public care," Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems, 2013, pp. 572-873, doi: 10.1109/CBMS.2013.6627881.

[60] Centralni zdravstveni informacijski sustav Republike Hrvatske (CEZIH), http://www.cezih.hr/eKarton.html, accessed December 2022.

[61] Portal zdravlja, https://portal.zdravlje.hr/portalzdravlja/login.html, accessed December 2022

[62] Javaid, M., Haleem, A., Rab, S., Singh, R. P., & Suman, R. (2021). Sensors for daily life: A review. Sensors International, Volume 2.

[63] Postolache, O. A., Mukhopadhyay, S. C., Jayasundera, K. P., & Swain, A. K. (2017). Sensors for Everyday Life. Springer.

[64] Dincer, Can, B., Richard, R., Estefanía, F.-A., M. Teresa, M., Arben, M., Güder, F. G. (2019). Disposable Sensors in Diagnostics, Food, and Environmental Monitoring. Advanced Materials, Volume 31.

[65] T. Mortenson, "What is a Sensor? Different Types of Sensors, Applications", August 2020, https://realpars.com/types-of-sensors/, accessed Dec 2022

[66] H. Baltes, O. Paul and O. Brand, "Micromachined thermally based CMOS microsensors," in Proceedings of the IEEE, vol. 86, no. 8, pp. 1660-1678, Aug. 1998, doi: 10.1109/5.704271.

[67] JCGM 200:2008 International vocabulary of metrology, International Bureau of Weights and Measures (BIPM), 2008

[68] [68] Bamford, Robert; Deibler, William (2003). ISO 9001: 2000 for Software and Systems Providers: An Engineering Approach (1st ed.). CRC-Press. ISBN 0-8493-2063-1, ISBN 978-0-8493-2063-7

[69] Evett, Steve & Tolk, Judy & Howell, Terry. (2006). Soil Profile Water Content Determination. Vadose Zone Journal. 5. 894. 10.2136/vzj2005.0149.

[70] Akyildiz, Ian & WY, Su & Sankarasubramaniam, Y. & Cayirci, E. (2002). Wireless Sensor Networks: A Survey. Computer Networks. 38. 393-422. 10.1016/S1389-1286(01)00302-4.

[71] O. Brdiczka et al., "Detecting human behavior models from multimodal observation in a smart home," IEEE Trans. on Automation Science and Engineering, vol. 6, pp. 588-597, 2009.

[72] R. M. Kwasnicki, S. Hettiaratchy, D. Jarchi, C. Nightingale, M. Wordsworth, J. Simmons, G. Z. Yang, A. Darzi. (2014, Jun), "Assessing functional mobility after lower limb reconstruction: A psychometric evaluation of a sensor-based mobility score," Annals of Surgery [Online]. pp. 1-7

[73] M. Howell Jones, A. Arcelus, R. Goubran, and F. Knoefel, "A pressure sensitive home environment," in IEEE HAVE, 2006, pp. 10-14.

[74] F.-T. Sun et al., "Activity-aware mental stress detection using physiological sensors," in MobiCASE, 2012, pp. 211-230.

[75] M. Borazio and K. Van Laerhoven, "Combining wearable and environmental sensing into an unobtrusive tool for long-term sleep studies," in Proc. ACM SIGHIT, 2012, pp. 71-80.

[76] J. Lanagan et al., "Utilising wearable and environmental sensors to identify the context of gait performance in the home," in Diverse, 2011, pp. 1-5.

[77] G.-Z. Yang, J. Andreu-Perez, X. Hu, and S. Thiemjarus, "Multisensor fusion," in Body sensor networks, 2nd ed., Germany: Springer, 2014, pp. 301-354.

[78] D. Yusufu, E. Magee, B. Gilmore, A. Mills, "Non-invasive, 3D printed, colourimetric, early wound-infection indicator", Chem. Commun., 2022,58, 439-442. DOI:10.1039/D1CC06147J

[79] Falcucci, T., Presley, K. F., Choi, J., Fizpatrick, V., Barry, J., Kishore, J., Ly, J. T., Grusenmeyer, T. A., Dalton, M. J., Kaplan, D. L., Degradable Silk-Based Subcutaneous Oxygen Sensors. Adv. Funct. Mater. 2022, 32, 2202020. DOI:10.1002/adfm.202202020

[80] Eun Hye Koh, Won-Chul Lee, Yeong-Jin Choi, Joung-Il Moon, Jinah Jang, Sung-Gyu Park, Jaebum Choo, Dong-Ho Kim, and Ho Sang Jung, „A Wearable Surface-Enhanced Raman Scattering Sensor for Label-Free Molecular Detection“, ACS Applied Materials & Interfaces 2021 13 (2), 3024-3032, DOI: 10.1021/acsami.0c18892

[81] S. Javaid, S. Zeadally, H. Fahim and B. He, "Medical Sensors and Their Integration in Wireless Body Area Networks for Pervasive Healthcare Delivery: A Review," in IEEE Sensors Journal, vol. 22, no. 5, pp. 3860-3877, 1 March1, 2022, doi: 10.1109/JSEN.2022.3141064.

[82] Facts & Factors. (2022). Insights on Global Wearable Technology Market Size & Share to Surpass USD 380.5 Billion by 2028, Exhibit a CAGR of 18.5% - Industry Analysis, Trends, Value, Growth, Opportunities, Segmentation, Outlook & Forecast Report. USA.

[83] Hatano, Y. (1997). Prevalence and use of pedometer. Res J Walk, 45-54.

[84] Tudor-Locke, C., Craig, C.L., Brown, W.J. et al. How many steps/day are enough? for adults. Int J Behav Nutr Phys Act 8, 79 (2011). DOI:10.1186/1479-5868-8-79

[85] Medical devices — Quality management systems — Requirements for regulatory purposes, ISO 13485:2016, Geneva, Switzerland: ISO.

[86] Medical electrical equipment — Part 2-56: Particular requirements for basic safety and essential performance of clinical thermometers for body temperature measurement, ISO 80601-2-56:2017, Geneva, Switzerland: ISO.

[87] General requirements for the competence of testing and calibration laboratories, ISO/IEC 17025:2017, Geneva, Switzerland: ISO.

[88] Bender, C. G., Hoffstot, J. C., Combs, B. T., Hooshangi, S., & Cappos, J. (2017). Measuring the fitness of fitness trackers. 2017 IEEE Sensors Applications Symposium (SAS), pp. 1-6. doi:10.1109/SAS.2017.7894077

[89] Andalibi, V., Honko, H., Christophe, F., & Viik, J. (2015). Data correction for seven activity trackers based on regression models. 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), (str. pp. 1592-1595). doi:10.1109/EMBC.2015.7318678

[90] Koren, A., Jurčević, M., & Prasad, R. (2021). Comparison of Data-Driven Models for Cleaning eHealth Sensor Data: Use Case on ECG Signal. Wireless personal communications, pp. 1501–1517.

[91] Gillinov S, Etiwy M, Wang R, Blackburn G, Phelan D, Gillinov AM, Houghtaling P, Javadikasgari H, Desai MY. Variable Accuracy of Wearable Heart Rate Monitors during Aerobic Exercise. Med Sci Sports Exerc. 2017 Aug;49(8):1697-1703. doi: 10.1249/MSS.0000000000001284. PMID: 28709155.

[92] Guy Hajj-Boutros, Marie-Anne Landry-Duval, Alain Steve Comtois, Gilles Gouspillou & Antony D. Karelis (2022) Wrist-worn devices for the measurement of heart rate and energy expenditure: A validation study for the Apple Watch 6, Polar Vantage V and Fitbit Sense, European Journal of Sport Science, DOI: 10.1080/17461391.2021.2023656

[93] Reddy RK, Pooni R, Zaharieva DP, Senf B, El Youssef J, Dassau E, Doyle Iii FJ, Clements MA, Rickels MR, Patton SR, Castle JR, Riddell MC, Jacobs PG. Accuracy of Wrist-Worn Activity Monitors During Common Daily Physical Activities and Types of Structured Exercise: Evaluation Study. JMIR Mhealth Uhealth. 2018 Dec 10;6(12):e10338. doi: 10.2196/10338. PMID: 30530451; PMCID: PMC6305876.

[94] Clinical trials. Accessed December 2022. https://clinicaltrials.gov/

[95] Fitbit Research Library. Fitabase. 2022. Accessed December 2022. https://www.fitabase.com/research-library/

[96] Evenson, K.R., Goto, M.M. & Furberg, R.D. Systematic review of the validity and reliability of consumer-wearable activity trackers. Int J Behav Nutr Phys Act 12, 159 (2015). doi: 10.1186/s12966-015-0314-1

[97] Lucisano, J. Y., Routh, T. L., Lin, J. T., & Gough, D. A. (2017). Glucose Monitoring in Individuals with Diabetes Using a Long-Term Implanted Sensor/Telemetry System and Model. IEEE Transactions on Biomedical Engineering, vol. 64, pp. 1982-1993.

[98] Padukone, G. S., & Devi, H. U. (2018). Tumor Markers for Cancer Detection using Optical Sensor. 2018 International Conference on Smart Systems and Inventive Technology (ICSSIT), (pp. 52-56).

[99] Divya, R., & Chinnaiyan, R. (2018). Reliable Smart Earplug Sensors for Monitoring Human Organs based on 5G Technology. 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), (str. pp. 687-690).

[100] Wang, F., & Liu, J. (2011). Networked wireless sensor data collection: Issues, challenges, and approaches. IEEE Communications Surveys & Tutorials, vol. 11, str. pp. 673–687.

[101] Dong, Y., Sun, L., Liu, D., Feng, M., & Miao, T. (2018). A Survey on Data Integrity Checking in Cloud. 2018 1st International Cognitive Cities Conference (IC3), (str. pp. 109-113).

[102] Hongyuan, W. (2019). An External Data Integrity Tracking and Verification System for Universal Stream Computing System Framework. 2019 21st International Conference on Advanced Communication Technology (ICACT), (str. pp. 32-37).

[103] Bhattacharjee, S., Salimitari, M., Chatterjee, M., Kwiat, K., & Kamhoua, C. (2017). Preserving Data Integrity in IoT Networks Under Opportunistic Data Manipulation. IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/ DataCom/CyberSciTech), (pp. pp. 446-453.).

[104] Huang, J. (2018). From Big Data to Knowledge: Issues of Provenance, Trust, and Scientific Computing Integrity. 2018 IEEE International Conference on Big Data, (str. pp. 2197-2205.).

[105] Ni, K.; Ramanathan, N.; Chehade, M.; Balzano, L.; Nair, S.; Zahedi, S.; Kohler, E.; Pottie, G.; Hansen, M.; Srivastava, M. (2009). Sensor network data fault types. ACM Transactions on Sensor Networks (TOSN), vol. 5, pp. 1-29.

[106] Parenreng, J. M., Kitagawa, A., & Andayani D., D. (2019). A Study of Limited Resources and Security Adaptation for Extreme Area in Wireless Sensor Networks. Journal of Physics: Conference Series.

[107] Parenreng, J. M., & Kitagawa, A. (2017). A Model of Security Adaptation for Limited Resources in Wireless Sensor Network. Journal of Computer and Communications, vol. 5, pp. 10-23.

[108] Parenreng, J. M., & Kitagawa, A. (2018). Techniques and Security Levels for Wireless Sensor Networks Based on the ARSy Framework. Sensors, vol. 18.

[109] Audéoud, H., & Heusse, M. (2018). Quick and Efficient Link Quality Estimation in Wireless Sensors Networks. 2018 14th Annual Conference on Wireless On-demand Network Systems and Services (WONS), (str. pp. 87-90).

[110] Karkouch, A., Mousannif, H., Moatassime, H. A., & Noel, T. (2016). Data quality in the internet of things: A state-of-the-art survey. Journal of Network and Computer Applications, vol. 73, pp. 57-81.

[111] Moghaddasi, H. (2016). A systemic biologic model for healthcare data quality. HIM-INTERCHANGE.

[112] Sion, D., Dooling, J., Glondys, B., Jones, D., Kadlec, L., Overgaard, M., Wendicke, A. (2015). Data Quality Management Model. Journal of AHIMA, vol. 86.

[113] Li, G., Peng, S., Wang, C., Niu, J., & Yuan, Y. (2019). An energy-efficient data collection scheme using denoising autoencoder in wireless sensor networks. Tsinghua Science and Technology, vol. 24, pp. 86-96.

[114] Fotiou, N., Siris, V. A., Mertzianis, A., & Polyzos, G. C. (2018). Smart IoT Data Collection. 2018 IEEE 19th International Symposium on "A World of Wireless, Mobile and Multimedia Networks" (WoWMoM), (str. pp. 588-599.).

[115] Schobel, J., Pryss, R., Schickler, M., & Reichert, M. (2016). Towards Flexible Mobile Data Collection in Healthcare. 2016 IEEE 29th International Symposium on Computer-Based Medical Systems (CBMS), (str. pp. 181-182).

[116] Islam, Z., Mamun, Q., & Rahman, G. (2014). Data Cleansing during Data Collection from Wireless Sensor Networks. Proceedings of the Twelfth Australasian Data Mining Conference (AusDM 2014).

[117] Dereszynski, E., & Diettrich, T. G. (2007). Probabilistic models for anomaly detection in remote sensor data streams. Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence (UAI2007), (str. pp. 75-82).

[118] Ramirez, G., Fuentes, O., & Tweedie, C. E. (2011). Assessing data quality in a sensor network for environmental monitoring. 2011 Annual Meeting of the North American Fuzzy Information Processing Society, (str. pp. 1-6).

[119] Cheng, H., Feng, D., Shi, X., & Chen, C. (2018). Data quality analysis and cleaning strategy for wireless sensor networks. EURASIP Journal on Wireless Communications and Networking.

[120] Rahman, M. G., Islam, M. Z., Bossomaier, T., & Gao, J. (2012). CAIRAD: A co-appearance-based analysis for Incorrect Records and Attribute-values Detection. The 2012 International Joint Conference on Neural Networks (IJCNN), (str. pp. 1-10).

[121] Rahman, M. G., & Islam, M. Z. (2014). FIMUS: A decision tree-based missing value imputation technique for data pre-processing. Proceedings of the Ninth Australasian Data Mining Conference 2014.

[122] Salem, O., Guerassimov, A., Mehaoua, A., Marcus, A., & Furht, B. (2013). Sensor fault and patient anomaly detection and classification in medical wireless sensor networks. 2013 IEEE International Conference on Communications (ICC), (str. pp. 4373-4378).

[123] Bruijn, B. d., Nguyen, T. A., Bucur, D., & Tei, K. (2016). Benchmark Datasets for Fault Detection and Classification in Sensor Data. SENSORNETS 2016 Proceedings of the 5th International Conference on Sensor Networks, (str. pp. 185-195.).

[124] Yang, D., Cheng, Y., Zhu, J., Xue, D., Abt, G., Ye, H., & Peng, Y. (2018). A Novel Adaptive Spectrum Noise Cancellation Approach for Enhancing Heartbeat Rate Monitoring in a Wearable Device. IEEE Access, vol. 6, pp. 8364-8375.

[125] Jauk, S., Kramer, D., & Leodolter, W. (2018). Cleansing and Imputation of Body Mass Index Data and Its Impact on a Machine Learning Based Prediction Model. Studies in Health Technology and Informatics, pp. 116-123.

[126] Nižetić-Kosović, I., Božić, A., Mastelić, T., & Ivanković, D. (2019). uilding Soft Sensors using Artificial Intelligence: Use Case on Daily Solar Radiation. 3rd International Conference on Smart and Sustainable Technologies.

[127] Yan, X., Xie, H., & Tong, W. (2011). A multiple linear regression data predicting method using correlation analysis for wireless sensor networks. Proceedings of 2011 Cross Strait Quad-Regional Radio Science and Wireless Technology Conference, (str. pp. 960-963).

[128] Qu, X., & Kim, H. J. (2014). Enhanced discriminant linear regression classification for face recognition. 2014 IEEE Ninth International Conference on Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP), (str. pp. 1-5).

[129] Tsang, S., Kao, B., Yip, K. Y., Ho, W., & Lee, S. D. (2011). Decision Trees for Uncertain Data. IEEE Transactions on Knowledge and Data Engineering, vol. 23, pp. 64-78.

[130] Sugiarto, B., & Sustika, R. (2016). Data classification for air quality on wireless sensor network monitoring system using decision tree algorithm. 2016 2nd International Conference on Science and Technology-Computer (ICST), (str. pp. 172-176).

[131] Ahmadi, A., Mitchell, E., Destelle, F., Gowing, M., O'Connor, N., Richter, C., & Moran, K. (2014). Automatic Activity Classification and Movement Assessment During a Sports Training Session Using Wearable Inertial Sensors. 2014 11th International Conference on Wearable and Implantable Body Sensor Networks, (pp. pp. 98-103).

[132] Rahman, M. J., & Morshed, B. I. (2019). Improving Accuracy of Inkjet Printed Core Body WRAP Temperature Sensor Using Random Forest Regression Implemented with an Android App. 2019 United States National Committee of URSI National Radio Science Meeting (USNC-URSI NRSM), (str. pp. 1-2).

[133] Al-Milli, N., & Almobaideen, W. (2019). Hybrid Neural Network to Impute Missing Data for IoT Applications. 2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT), (str. pp. 121-125).

[134] Ahmadi, H., & Bouallegue, R. (2015). Comparative study of learning-based localization algorithms for Wireless Sensor Networks: Support Vector regression, Neural Network and Naïve Bayes. 2015 International Wireless Communications and Mobile Computing Conference (IWCMC), (str. pp. 1554-1558).

[135] Yang, D., Chhatre, N., Campi, F., & Menon, C. (2016). Feasibility of Support Vector Machine gesture classification on a wearable embedded device. 2016 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE), (str. pp. 1-4).

[136] Banos, Garcia, Holgado, Damas, Pomares, Rojas, Villalonga. (2014). mHealthDroid: a novel framework for agile development of mobile health applications. Proceedings of the 6th International Work-conference on Ambient Assisted Living an Active Aging (IWAAL 2014).

[137] Banos, Villalonga., Garcia., Saez., Damas., Holgado, Rojas. (2015). Design, implementation and validation of a novel open framework for agile development of mobile health applications. BioMedical Engineering OnLine, vol. 14, pp. 1-20.

[138] Rahman, G., & Islam, Z. (2013). Missing Value Imputation Using Decision Trees and Decision Forests by Splitting and Merging Records: Two Novel Techniques. Knowledge-Based Systems, vol. 53, pp. 51-65.

[139] Rennie, K. L., Hemingway, H., Kumari, M., Brunner, E., Malik, M., & Marmot, M. (2003). Effects of Moderate and Vigorous Physical Activity on Heart Rate Variability in a British Study of Civil Servants. American Journal of Epidemiology, vol. 158, pp. 135–143.

[140] Zhang, Z., Pi, Z., & Liu, B. (2015). TROIKA: A General Framework for Heart Rate Monitoring Using Wrist-Type Photoplethysmographic Signals During Intensive Physical Exercise. IEEE Transactions on Biomedical Engineering, vol. 62, pp. 522-531.

[141] Long, X., Yin, B., & Aarts, R. M. (2009). Single-accelerometer-based daily physical activity classification. 2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, (pp. pp. 6107-6110).

[142] Gyllensten, I. C., & Bonomi, A. G. (2011). Identifying Types of Physical Activity with a Single Accelerometer: Evaluating Laboratory-trained Algorithms in Daily Life. IEEE Transactions on Biomedical Engineering, vol. 58, pp. 2656-2663.

[143] Al-Fatlawi, A. H., Fatlawi, H. K., & Ling, S. H. (2017). Recognition physical activities with optimal number of wearable sensors using data mining algorithms and deep belief network. 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), (pp. pp. 2871-2874).

[144] Awais, M., Palmerini, L., & Chiari, L. (2016). Physical activity classification using body-worn inertial sensors in a multi-sensor setup. 2016 IEEE 2nd International Forum on Research and Technologies for Society and Industry Leveraging a better tomorrow (RTSI), (str. pp. 1-4).

[145] Fahim, M., Khattak, A. M., Aleem, S., & Katheeri, H. A. (2017). Physical activity recognizer based on multimodal sensors in smartphones for ubiquitous-lifecare services. 2017 IEEE AFRICON, (str. pp. 524-529).

[146] Li, P., Wang, Y., Tian, Y., Zhou, T., & Li, J. (2017). An Automatic User-Adapted Physical Activity Classification Method Using Smartphones. IEEE Transactions on Biomedical Engineering, vol. 64, pp. 706-714.

[147] Koren, A., Jurčević, M., & Huljenić, D. (2019). Requirements and Challenges in Integration of Aggregated Personal Health Data for Inclusion into Formal Electronic Health Records (EHR). 2019 2nd International Colloquium on Smart Grid Metrology (SMAGRIMET), (str. pp. 1-5). doi:10.23919/SMAGRIMET.2019.8720389

[148] Thambawita, V., Hicks, S., Borgli, H., Pettersen, S. A., Stensland, H. K., Jha, D., & Johansen, D. (2020). Pmdata: A Sports Logging Dataset. OSF Preprints.

[149] Koren, A. (2021). OxyBeat Dataset. Accessed December 2021, retrieved from https://github.com/korenana/oxybeat-dataset

[150] Fitbit Reports. (2020). Third Quarter Results. Retrieved from https://investor.fitbit.com/press/press-releases/press-release-details/2020/Fitbit-Reports-Third-Quarter-Results-for-the-Three-Months-Ended-October-3-2020/default.aspx

[151] Mishra, A., Nieto, A., & Kitsiou, S. (2017). Systematic Review of mHealth Interventions Involving Fitbit Activity Tracking Devices. 2017 IEEE International Conference on Healthcare Informatics (ICHI), (str. pp. 455-455).

[152] US National Library of Medicine. (3rd. October 2021). ClinicalTrials.gov.

[153] Vlatakis, G., Andersson, L., & Müller, R. (1993). Drug assay using antibody mimics made by molecular imprinting. Nature, 645–647.

[154] Rosenberg, Kadokura, Bouldin, Miyawaki, Higano, & Hartzler. (2016). Acceptability of Fitbit for physical activity tracking within clinical care among men with prostate cancer. AMIA Annu Symp Proc. 2016, (pp. pp. 1050–1059).

[155] FHIR Heart Rate Profile. (2018, August 11th). Retrieved from http://hl7.org/fhir/StructureDefinition/heartrate

[156] FHIR StructureDefinition: VitalSigns. (2016, March 25th). Retrieved from http://hl7.org/fhir/StructureDefinition/vitalsigns

[157] FHIR Body Temperature Profile. (2018, August 11th). Retrieved from http://hl7.org/fhir/StructureDefinition/bodytemp

[158] FHIR Oxygen Saturation Profile. (2018, October 23rd). Retrieved from http://hl7.org/fhir/StructureDefinition/oxygensat

[159] Rahman, M., Carbinar, B. & Banik, M. (2013). Fit and Vulnerable: Attacks and Defenses for a Health Monitoring Device. 34th IEEE Symposium on Security and Privacy (IEEE S\&P).

[160] Trackers: Fit for Health but Unfit for Security and Privacy. 2017 IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE), (str. pp. 19-24). doi:10.1109/CHASE.2017.54

[161] Shekar, A. R. (2019). Preventing Data Manipulation and Enhancing the Security of data in Fitness Mobile Application. 2019 International Conference on Smart Systems and Inventive Technology (ICSSIT), (str. pp. 740-745). doi:10.1109/ICSSIT46314.2019.8987892

[162] Lotfy, K., & Hale, M. L. (2016). Assessing Pairing and Data Exchange Mechanism Security in the Wearable Internet of Things. 2016 IEEE International Conference on Mobile Services (MS), (str. pp. 25-32).

[163] Hale, M. L., Ellis, D., Gamble, R., Waler, C., & Lin, J. (2015). Secu Wear: An Open Source, Multi-component Hardware/Software Platform for Exploring Wearable Security. 2015 IEEE International Conference on Mobile Services, (str. pp. 97-104).

[164] Anwar, S., Anwar, D., & Abdulla, S. (2020). Security Evaluation of Android Mobile Healthcare and Fitness Applications. 2020 International Conference on Electrical, Communication, and Computer Engineering (ICECCE), (str. pp. 1-6).

[165] Rahman, M., Carbunar, B., & Topkara, U. (2014). SensCrypt: A Secure Protocol for Managing Low Power Fitness Trackers. 2014 IEEE 22nd International Conference on Network Protocols, (str. pp. 191-196).

[166] Saha, R., Sarkar, S., & Datta, S. K. (2017). Balancing security & sharing of fitness trackers' data. 2017 1st International Conference on Electronics, Materials Engineering and Nanotechnology (IEMENTech), (str. pp. 1-6).

[167] Meinert, D. B., Peterson, D. K., Criswell, J. R., & Crossland, M. D. (2006). Privacy policy statements and consumer willingness to provide personal information. Journal of Electronic Commerce in Organizations (JECO), pp. 1-17.

[168] Schomakers, E., Lidynia, C., & Ziefle, M. (2019). Listen to My Heart? How Privacy Concerns Shape Users' Acceptance of e-Health Technologies. 2019 International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob), (str. pp. 306-311). doi:10.1109/WiMOB.2019.8923448

[169] Rahman, M. S. (2019). Does Privacy Matters When We are Sick? An Extended Privacy Calculus Model for Healthcare Technology Adoption Behavior. 2019 10th International Conference on Information and Communication Systems (ICICS), (str. pp. 41-46). doi:10.1109/IACS.2019.8809175

[170] Siriwardhana, Y., Porambage, P., Liyanage, M., & Ylianttila, M. (2021). AI and 6G Security: Opportunities and Challenges. 2021 Joint European Conference on Networks and Communications & 6G Summit (EuCNC/6G Summit), (str. pp. 616-621). doi:10.1109/EuCNC/6GSummit51104.2021.9482503

[171] Zuhair, M., Patel, F., Navapara, D., Bhattacharya, P., & Saraswat, D. (2021). BloCoV6: A blockchain-based 6G-assisted UAV contact tracing scheme for COVID-19 pandemic. 2021 2nd International Conference on Intelligent Engineering and Management (ICIEM), (str. pp. 271-276). doi:10.1109/ICIEM51511.2021.9445332

[172] Pribadi, I. L., & Suryanegara, M. (2017). Regulatory recommendations for IoT smart-health care services by using privacy impact assessment (PIA). 2017 15th International Conference on Quality in Research (QiR), International Symposium on Electrical and Computer Engineering, (str. pp. 491-496). doi:10.1109/QIR.2017.8168535

[173] A. Koren, M. Jurčević and R. Prasad, "Semantic Constraints Specification and Schematron-Based Validation for Internet of Medical Things' Data," in IEEE Access, vol. 10, pp. 65658-65670, 2022, doi: 10.1109/ACCESS.2022.3182486.

[174] Coons, Creech, & Jones. (1941). Immunological Properties of an Antibody Containing a Fluorescent Group. Proceedings of the Society for Experimental Biology and Medicine, 200-202.

[175] Takátsy, G. (2003). The Use of Spiral Loops in Serological and Virological Micro-Methods. Acta Microbiologica et Immunologica Hungarica, 369-383.

[176] Jacobson, B., & Mackay, R. (1967). A pH-endoradiosonde. Lancet.

[177] Rosalyn, S. Y., & Berson, S. A. (1959). Assay of Plasma Insulin in Human Subjects by Immunological Methods. Nature, 1648-1649.

[178] Clark, L., & Lyons, C. (1962). Electrode systems for continuous monitoring in cardiovascular surgery. Annals of the New York Academy of Sciences.

[179] Avrameas, S., & Uriel, J. (1966). Method of antigen and antibody labelling with enzymes and its immunodiffusion application. Comptes Rendus Hebdomadaires des Seances de l Academie des Sciences, 2543-2545.

[180] Anderson NG, (1969). Analytical techniques for cell fractions. XII. A multiple-cuvet rotor for a new microanalytical system. Analytical Biochemistry, 545-562.

[181] Engvall, E., & Perlmann, P. (1971). Enzyme-linked immunosorbent assay (ELISA) - Quantitative assay of immunoglobulin G. Immunochemistry, 871-874.

[182] Bergveld, P. (1972). Development, operation, and application of the ion-sensitive field-effect transistor as a tool for electrophysiology. IEEE Transactions on Biomedical Engineering, 342-351.

[183] Köhler, G., & Milstein, C. (1975). Continuous cultures of fused cells secreting antibody of predefined specificity. Nature, 495-497.

[184] Mulle, R., & Howe, R. T. (1983). Polycrystalline Silicon Micromechanical Beams. Journal of The Electrochemical Society.

[185] Liedberg, B., Nylander, C., & Lunström, I. (1983). Surface plasmon resonance for gas detection and biosensing. Sensors and Actuators, Volume 4, 299-304.

[186] Cass, Davis, Francis, Hill, Aston, Higgins, Turner. (1984). Ferrocene-Mediated Enzyme Electrode for Amperometric Determination of Glucose. Analytical Chemistry, 667-671.

[187] Ekins, R. (1989). Multi-analyte immunoassay. Journal of Pharmaceutical and Biomedical Analysis, 155-168.

[188] Manz, A., Graber, N., & Widmer, H. (1990). Miniaturized total chemical analysis systems: A novel concept for chemical sensing. Sensors and Actuators B: Chemical, Volume 1, Issues 1–6, 244-248.

[189] Ellington, A., & Szostak, J. (1990). In vitro selection of RNA molecules that bind specific ligands. Nature, 818-822.

[190] Vlatakis, G., Andersson, L., Müller, R. et al. Drug assay using antibody mimics made by molecular imprinting. Nature 361, 645–647 (1993). https://doi.org/10.1038/361645a0

[191] Kim, E., Xia, Y., & Whitesides, G. (1995). Polymer microstructures formed by moulding in capillaries. Nature, 581-584.

[192] Jobst, G., Moser, I., Svasek, P., Varahram, M., Trajanoski, Z., Wach, P., Urban, G. (1997). Mass producible miniaturized flow through a device with a biosensor array. Sensors & Actuators: B. Chemical, 121-125.

[193] Vogelstein, B., & Kinzler, K. (1999). Digital PCR. The Proceedings of the National Academy of Sciences (PNAS), 9236-9241.

[194] Martinez, A., Phillips, S., Butte, M., & Whitesides, G. (2007). Patterned paper as a platform for inexpensive, low-volume, portable bioassays. Angewandte Chemie International Edition English, 1318-1320.

[195] Huh, D., Matthews, B., Mammoto, A., Montoya-Zavala, M., Hsin, H., & Ingber, D. (2010). Reconstituting organ-level lung functions on a chip. Science, 1662-1668.

[196] Kim, DH; Lu, N; Ma, R; Kim, YS; Kim, RH; Wang, S; Wu, J; Won, SM; Tao, H; Islam, A; Yu, KJ; Kim, TI; Chowdhury, R,; Ying, M,; Xu, L,; Li, M; Chung, HJ; Keum, H; McCormick, M; Liu, P; Zhang, YW; Omenetto, FG; Huang, Y; Coleman, T; Rogers, JA, (2011). Epidermal electronics. Science, 838-843.

[197] Morales-Narváez, E., Baptista-Pires, L., Zamora-Gálvez, A., & Merkoçi, A. (2016). Graphene-Based Biosensors: Going Simple. Advanced Materials.

[198] Gootenberg, J., Abudayyeh, O., Lee, J., Essletzbichler, P., Dy, A., Joung, J., Zhang, F. (2017). Nucleic acid detection with CRISPR-Cas13a/C2c2. Science, 438-442.

[199] Schliemann T., Danielsen C., Virtanen T., Vuokko R., Hardardottir G. A., Alsaker M. A., Asnes B., Eklöf N., Ericsson E., "eHealth standardisation in the Nordic countries", Nordic Council of Ministers, 2019.

[200] Electronic Health Record (EHR) Standards for India, e-Health Division, Department of Health & Family Welfare, Ministry of Health & Family Welfare, Government of India, 2016.

[201] Schematron implementation, Schematron.com/home/implementation.html, accessed 5.3.2023.

[202] Oracle® Fusion Middleware, Introduction to API Gateway OAuth 2.0., https://docs.oracle.com/cd/E50612_01/doc.11122/oauth_guide/content/oauth_intro.html, accessed 21.3.2023.

[203] Bassett DR Jr, Rowlands A, Trost SG. Calibration and validation of wearable monitors. Med Sci Sports Exerc. 2012 Jan;44(1 Suppl 1): S32-8. doi: 10.1249/MSS.0b013e3182399cf7.

[204] Straiton N, Alharbi M, Bauman A, Neubeck L, Gullick J, Bhindi R, Gallagher R. The validity and reliability of consumer-grade activity trackers in older, community-dwelling adults: A systematic review. Maturitas. 2018 Jun; 112:85-93. doi: 10.1016/j.maturitas.2018.03.016.

[205] E. Allegretti, E. Sibilano, A. Di Nisio, A. M. L. Lanzolla and M. Spadavecchia, "Assessment and Calibration of Wearable Heart Rate Sensors Using a Fully Automated System," 2020 IEEE International Symposium on Medical Measurements and Applications (MeMeA), Bari, Italy, 2020, pp. 1-6, doi: 10.1109/MeMeA49120.2020.9137329.

[206] Devin Warner Miller, David Rich Miller, Jeffrey Michael Lee, "Calibration of a wearable medical device". U.S. Patent No. 20150289820, U.S. Patent and Trademark Office.

[207] Meru Adagouda PATIL, Nagaraju Bussa, Prasad RAGHOTHAM VENKAT, "Methods and apparatus for calibrating a medical monitoring device", U.S. Patent No. 2017102521, U.S. Patent and Trademark Office.

Abbreviations (Glossary)

ADL	Activities of Daily Living
AES	Advanced Encryption Standard
AHIMA	American Health Information Management Association
AI	Artificial Intelligence
ANSI	American National Standards Institute
API	Application Programming Interface
AQL	Archetype Query Language
ARDS	Acute Respiratory Distress Syndrome
ASNC	Adaptive Spectrum Noise Cancellation
BIPM	International Bureau of Weights and Measures
CAGR	Compound Annual Growth Rate
CC	Cloud Computing
CCD	Continuity of Care Document
CCR	Continuity of Care Record
CDA	Clinical Document Architecture
CDU	Cough Detecting Unit
CEZIH	Centralni zdravstveni informacijski sustav Republike Hrvatske (Information Health System of the Republic of Croatia)
COVID-19	Coronavirus Disease 2019
CV	Coefficient of Variation
E2E	End-to-End
ECG	Electrocardiogram
EE	Energy Expenditure
EHR	Electronic Health Record
EMR	Electronic Medical Record
EU	European Union
FHIR	Fast Healthcare Interoperability Resources
GDPR	General Data Protection Regulation
GUI	Graphical User Interface
HIPAA	Health Insurance Portability and Accountability Act
HIS	Health Information System
HL7	Health Level Seven

HTTP	Hypertext Transfer Protocol
HZZO	Hrvatski zavod za zdravstveno osiguranje (Croatian Health Insurance Fund)
IEEE	Institute of Electrical and Electronics Engineers
IHE	Integrating the Healthcare Enterprise
ISO	International Organization for Standardization
IoMT	Internet of Medical Things
IoT	Internet of Things
JSON	JavaScript Object Notation
LOB	Limit of Blank
LOD	Limit of Detection
LOQ	Limit of Quantification
LQI	Link Quality Indicator
MLR	Multiple Linear Regression
mURLLC	massive Ultra-Reliable Low-Latency Communication
PCHR	Personally Controlled Health Record
PHR	Personal Health Record
PHRS	Personal Health Record System
PING	Personal Internetworked Notary and Guardian
POMR	Problem-Oriented Medical Record
PPE	Personal Protective Equipment
PPG	Photoplethysmography
PSI	Pounds per Square Inch
QOD	Quality of Data
QOI	Quality of Information
QOS	Quality of Service
RBAC	Role-Based Access Control
REST	REpresentational State Transfer
RIM	Reference Information Model
RIS	Reconfigurable Intelligent Surfaces
RMSE	Root Mean Square Error
RRMSE	Relative Root Mean Square Error
RTD	Resistance Temperature Detector
SERS	Surface-Enhanced Raman Scattering

SES	Self Encryption System
SHA	Secure Hash Algorithm
SMART	Substitutable Medical Applications, Reusable Technologies
SOA	Service-Oriented Architecture
SQL	Structured Query Language
SSL	Secure Sockets Layer
SVM	Support Vector Machine
TR	Technical Report
TSU	Temperature Sensing Unit
WHO	World Health Organization
WSN	Wireless Sensor Network
XML	eXtensible Markup Language
XSLT	Extensible Stylesheet Language Transformations

Biography

Ana Koren finished her bachelor's degree in Telecommunication and Information Technology in 2012 and master's degree in Information and Communication Technology in 2014 at University of Zagreb. She worked as research associate at Faculty of Electrical Engineering and Computing, University of Zagreb on EU FP7 project eWALL: for Active Long Living and has been a visiting researcher at various universities, such as Universidad de Zaragoza, in Zaragoza, Spain and Universidad Nacional de Colombia, in Bogota, Colombia. She is currently working as a software developer at Ericsson Nikola Tesla.

List of Pulications

- Ana Koren, Marko Jurčević, Ramjee Prasad, "Comparison of Data-Driven Models for Cleaning eHealth Sensor Data: Use Case on ECG Signal", Wireless Personal Communications, vol.114, no.2, pp.1501, 2020., doi: 10.1007/s11277-020-07435-7
- Ana Koren, Marko Jurčević, Ramjee Prasad, "Semantic Constraints Specification and Schematron-Based Validation for Internet of Medical Things' Data", IEEE Access, vol.10, pp. 65658-65670, 2022, doi: 10.1109/ACCESS.2022.3182486.
- Ana Koren and Ramjee Prasad, "IoT Health Data in Electronic Health Records (EHR): Security and Privacy Issues in Era of 6G," in Journal of ICT Standardization, vol. 10, no. 1, pp. 63-84, 2022, doi: 10.13052/jicts2245-800X.1014.

- Ana Koren, Marko Jurčević and Darko Huljenić, "Requirements and Challenges in Integration of Aggregated Personal Health Data for Inclusion into Formal Electronic Health Records (EHR)," 2019 2nd International Colloquium on Smart Grid Metrology (SMAGRIMET), Split, Croatia, 2019, pp. 1-5, doi: 10.23919/SMAGRIMET.2019.8720389
- Ana Koren, Ramjee Prasad, "Personal Wireless Data in Formal Electronic Health Records: Future Potential of Internet of Medical Things Data", 2020 23rd International Symposium on Wireless Personal Multimedia Communications (WPMC), pp. 1-4, doi: 10.1109/WPMC50192.2020.9309482
- Ana Koren, Marko Jurčević, "Concept-Level Model of Integrated Syntax and Semantic Validation for Internet of Medical Things Data", 2021 IEEE 15th International Conference on Semantic Computing (ICSC), Laguna Hills, CA, USA, 2021, pp. 207-210, doi: 10.1109/ICSC50631.2021.00044

- Ana Koren, Ramjee Prasad, "Internet of Things: Shaping Healthcare during COVID-19 Pandemic," 2021 24th International Symposium on Wireless Personal Multimedia Communications (WPMC), Okayama, Japan, 2021, pp. 1-6, doi: 10.1109/WPMC52694.2021.9700472
- Ana Koren, Ramjee Prasad, "Standardization of Third-party Data in Electronic Health Records", 2022 25th International Symposium on Wireless Personal Multimedia Communications (WPMC), pp. 453-458, doi: 10.1109/WPMC55625.2022.10014929

Životopis

Ana Koren završila je diplomski studij Telekomunikacijskih i informacijskih tehnologija 2012. godine i magisterij Informacijskih i komunikacijskih tehnologija 2014. godine na Sveučilištu u Zagrebu. Radila je kao znanstvena suradnica na Fakultetu elektrotehnike i računarstva Sveučilišta u Zagrebu na EU FP7 projektu eWALL: for Active Long Living te je bila gostujući istraživač na raznim sveučilištima, kao što su Universidad de Zaragoza, u Zaragozi, Španjolska i Universidad Nacional de Colombia, u Bogoti, Kolumbija. Trenutno radi kao programer u Ericssonu Nikoli Tesli.