

Intelligent Systems Reference Library 258

Chee-Peng Lim · Ashlesha Vaidya ·  
Nikhil Jain · Margarita N. Favorskaya ·  
Lakhmi C. Jain *Editors*

# Advances in Intelligent Healthcare Delivery and Management


Research Papers in Honour of Professor  
Maria Virvou for Invaluable Contributions

 Springer

# **Intelligent Systems Reference Library**

Volume 258

## **Series Editors**

Janusz Kacprzyk , Polish Academy of Sciences, Warsaw, Poland

Lakhmi C. Jain, KES International, Shoreham-by-Sea, UK

The aim of this series is to publish a Reference Library, including novel advances and developments in all aspects of Intelligent Systems in an easily accessible and well structured form. The series includes reference works, handbooks, compendia, textbooks, well-structured monographs, dictionaries, and encyclopedias. It contains well integrated knowledge and current information in the field of Intelligent Systems. The series covers the theory, applications, and design methods of Intelligent Systems. Virtually all disciplines such as engineering, computer science, avionics, business, e-commerce, environment, healthcare, physics and life science are included. The list of topics spans all the areas of modern intelligent systems such as: Ambient intelligence, Computational intelligence, Social intelligence, Computational neuroscience, Artificial life, Virtual society, Cognitive systems, DNA and immunity-based systems, e-Learning and teaching, Human-centred computing and Machine ethics, Intelligent control, Intelligent data analysis, Knowledge-based paradigms, Knowledge management, Intelligent agents, Intelligent decision making, Intelligent network security, Interactive entertainment, Learning paradigms, Recommender systems, Robotics and Mechatronics including human-machine teaming, Self-organizing and adaptive systems, Soft computing including Neural systems, Fuzzy systems, Evolutionary computing and the Fusion of these paradigms, Perception and Vision, Web intelligence and Multimedia.

Indexed by SCOPUS, DBLP, zbMATH, SCImago.

All books published in the series are submitted for consideration in Web of Science.

Chee-Peng Lim · Ashlesha Vaidya · Nikhil Jain ·  
Margarita N. Favorskaya · Lakhmi C. Jain  
Editors

# Advances in Intelligent Healthcare Delivery and Management

Research Papers in Honour of Professor  
Maria Virvou for Invaluable Contributions

 Springer

*Editors*

Chee-Peng Lim  
Department of Computing Technologies  
Swinburne University of Technology  
Hawthorn, VIC, Australia

Ashlesha Vaidya  
Flinders Medical Centre  
Bedford Park  
Adelaide, SA, Australia

Nikhil Jain  
The Permanente Medical Group, Inc.  
Northern California Region  
Fairfield, CA, USA

Margarita N. Favorskaya  
Reshetnev Siberian State University  
of Science and Technology  
Krasnoyarsk, Russia

Lakhmi C. Jain  
University of Piraeus  
Athens, Greece

ISSN 1868-4394

ISSN 1868-4408 (electronic)

Intelligent Systems Reference Library

ISBN 978-3-031-65429-9

ISBN 978-3-031-65430-5 (eBook)

<https://doi.org/10.1007/978-3-031-65430-5>

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2024

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG  
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

If disposing of this product, please recycle the paper.

# Preface

In this fast-moving digital era, the advent of Artificial Intelligence (AI) to innovate and improve healthcare services marks a pivotal moment. This volume brings together leading experts to illuminate this transformative journey, from electronic health records, robotics, and AI in healthcare, to machine learning-based decision support for patient self-triage and appointment scheduling. Each chapter offers a glimpse into the intersection between intelligence and healthcare. It is envisaged that this volume could provide readers with useful knowledge on how intelligent systems enhance patient care, optimize resources, and revolutionize healthcare delivery and management. A description of each chapter collected in this volume is as follows.

Maghool et al. discuss the importance of advanced analytics in utilizing Electronic Health Records (EHRs) for improving healthcare services. The use of distributed computing and AI in smart healthcare is examined, presenting practical use cases to enhance patient care. A workflow using the SMART BEAR (a big data platform with evidence-based personalized support for healthy and independent living at home) infrastructure is devised to standardize and improve data quality. Its usefulness in predicting future health conditions like cardiovascular disease and mild depression is assessed.

Wiczorek examines the critical role of usability and user experience, including emotion, trust, and ethical consideration, pertaining to AI applications in healthcare. The importance of following human-centered design principles to meet user needs is emphasized. While medical professionals are the primary users of AI-based healthcare tools, involving patients in the development process is crucial, particularly in contexts like chronic condition prevention and management. The study highlights older persons as a significant user group with specific needs and limitations, advocating for their inclusion in the human-centered design process in the design and development of AI-based healthcare systems.

Gino et al. study the impact of the COVID-19 pandemic on telehealth, focusing on the role of tele-education and tele-simulation in healthcare. The emergence and effectiveness of tele-simulation for safe clinical skills training are examined. A research and innovation laboratory, i.e., maxSIMhealth lab, for addressing healthcare training

challenges through innovative simulation solutions is introduced. The study underscores the increased integration of technology in healthcare education, driven by the need for swift adaptation during the pandemic.

Luca et al. address the challenges associated with storing personal medical data in a globalized context, covering existing and potential AI-powered systems. Accessing large databases presents difficulties owing to regulations (e.g., the General Data Protection Regulation). Despite the challenges, it is important to explore different aspects and possibilities in storing individual health information to advance medical decision support systems. Indeed, delivering timely, accurate, and personalized information to healthcare professionals in decision-making processes shows promise in enhancing healthcare outcomes.

Cuza et al. investigate the crucial issue of determining the minimum data requirement for training and testing AI models. A case study focusing on semantic segmentation of radiology images is presented. A combination of theoretical insights with experimental results offers comprehensive guidance for various phases of model development. The study encompasses both supervised and zero-shot segmentation approaches, which include the “Segment Anything Model”, providing a holistic understanding of the subject matter.

Kolpashchikov et al. delve into the growing significance of robots and AI in healthcare, playing diverse roles across different fields and delivering benefits to both patients and healthcare professionals. An overview on how robots and AI interact within the medical field is presented. Robots are useful in conducting precise surgical procedures and for helping disabled individuals and caregivers. Meanwhile, AI enhances diagnostic accuracy, offers mental support, and facilitates medical education. By integrating AI, the capabilities of medical robotics can be greatly enhanced.

Friday et al. strive to answer the question: “How can a network of aged and community care living labs provide the benefits of embedded technology innovation while overcoming the limitations of translation and scaling in the Australian context?” Through co-design with international and local living lab experts and aged care providers, a purpose-built network and process model is developed to mitigate issues associated with place-based approaches. The findings contribute towards addressing the growing importance of technology-supported aged and community care solutions due to demographic transitions, particularly in diverse settings like rural and urban areas.

Campolina and Abe analyze the critical role of patients and healthcare experts in Health Technology Assessment (HTA). Advocating a Paraconsistent Expert-Based Multi-Criteria Decision Analysis approach, and collaborative methodologies with experts and HTA organizations are discussed. The significance of incorporating expert perspectives in HTA, including patients, is highlighted. Paraconsistent approaches, as exemplified by the Paraconsistent value framework, offer a robust methodology, enhancing decision legitimacy and transparency in HTA.

Belciug & Iliescu explore two methods for determining the correct view plane in examining fetal morphology ultrasound videos, i.e., a probabilistic approach and a peer pressure approach. Convolutional neural networks (CNNs) are trained to identify

fetal abdomen view planes from image scans. A statistical analysis is applied to select the appropriate method for analyzing ultrasound videos, considering challenges such as fetal movement and overlapping view planes. Both methods perform well, with the results validated by expert sonographers.

Li et al. evaluate the Liver Imaging Reporting and Data System (LI-RADS), which standardizes liver imaging terminology, interpretation, and reporting, for undertaking liver tumor malignancy. An intelligent LI-RADS framework utilizing multi-task CNNs is developed. The framework incorporates four distinct branches to extract and analyze specific pathological features. The new multi-task deep learning framework is able to achieve reliable and high-precision LI-RADS classification results.

Ng et al. assess the complexity of healthcare logistics, addressing the challenge of efficient last-mile delivery in ensuring timely delivery of drugs, medical supplies, and others, for improving healthcare services. The Tabu Search (TS) algorithm is utilized as a solution for optimizing delivery routes in healthcare logistics. Following a rigorous systems development life cycle, the developed software tool is evaluated using three daily logistics scenarios, demonstrating its effectiveness in real-world applications.

Ong et al. develop machine learning (ML) models for use as a self-triage decision support tool. The tool predicts and assigns medical specialists to patients based on patient symptoms. Specifically, ML is devised by following several steps, including data pre-processing, feature selection, and hyper-parameter tuning. The predictions and implications of ML-based models are analyzed through a case study involving patient self-triage and appointment scheduling.

We wish to express our heartfelt gratitude to the esteemed authors whose contributions have enriched this volume with inspiring research and profound insights. We also would like to express our sincere appreciation to the reviewers for their efforts in shaping this volume into a comprehensive resource. The dedication and expertise of all parties involved have illuminated the path towards leveraging intelligent methodologies to deliver better healthcare services for our society.

Hawthorn, Australia  
Adelaide, Australia  
Fairfield, USA  
Krasnoyarsk, Russian Federation  
Athens, Greece

Chee-Peng Lim  
Ashlesha Vaidya  
Nikhil Jain  
Margarita N. Favorskaya  
Lakhmi C. Jain



# Contents

<b>1</b>	<b>Technologies and Strategies for Continuous Learning through Electronic Health Records Data</b> .....	<b>1</b>
	Samira Maghool, Valerio Bellandi, and Paolo Ceravolo	
<b>2</b>	<b>Human-Centered Design of AI in Healthcare and the Role of Older Patients</b> .....	<b>37</b>
	Rebecca Wiczorek	
<b>3</b>	<b>Telehealth: A Game Changer in Global Health Professions Education and Patient Care</b> .....	<b>49</b>
	Bruno Gino, Sandy Abdo, Bill Kapralos, and Adam Dubrowski	
<b>4</b>	<b>Individual Health Data Storage for Diagnosis and Decision Support Systems—Considerations on Colonoscopy Assessment</b> .....	<b>69</b>
	Mihaela Luca, Adrian Ciobanu, and Vlad Constantin Crăciun	
<b>5</b>	<b>Robotics and Artificial Intelligence in Healthcare</b> .....	<b>93</b>
	Dmitrii Kolpashchikov, Olga Gerget, and Roman Meshcheryakov	
<b>6</b>	<b>Sample Size for Training and Testing: Segment Anything Models and Supervised Approaches</b> .....	<b>107</b>
	Daniela Cuza, Carlo Fantozzi, Loris Nanni, Daniel Fusaro, Gustavo Zanoni Felipe, and Sheryl Brahnham	
<b>7</b>	<b>Building a Living Lab Network to Support Technology Innovation Within the Australian Aged Care Sector</b> .....	<b>147</b>
	Gareth Priday, Sonja Pedell, Anne Livingstone, George Margelis, and Georgina Gould	
<b>8</b>	<b>Patients and Healthcare Professionals Participation in Health Technology Assessment</b> .....	<b>171</b>
	Alessandro Gonçalves Campolina and Jair Minoro Abe	

**9 From Ultrasound Image Classification to Ultrasound Video Classification Approaches** ..... 189  
Smaranda Belciug and Dominic Gabriel Iliescu

**10 Enhanced Liver Imaging Reporting and Data System (LI-RADS) Through Multi-task Convolutional Neural Networks** ..... 201  
Yinhao Li, Qingqing Chen, Rahul Kumar Jain, Fang Wang, Hongjie Hu, Lanfen Lin, and Yen-Wei Chen

**11 Metaheuristic Tabu Search for Vehicle Scheduling: A Case Study of Healthcare Logistics** ..... 221  
Xin Ju Ng, Yi Wen Kerk, Ting Yee Lim, and Choo Jun Tan

**12 A Machine Learning Based Decision Support System for Healthcare Triage Applications** ..... 237  
Yi Chen Ong, Sim Ee Kee, Koh Kiong Chai, Ting Yee Lim, and Choo Jun Tan

# Chapter 1

## Technologies and Strategies for Continuous Learning through Electronic Health Records Data



Samira Maghool, Valerio Bellandi, and Paolo Ceravolo

**Abstract** Achieving a comprehensive view of a patient's health using data from Electronic Health Record systems requires the use of advanced analytics. However, effectively managing and curating this data requires carefully designed workflows. While digitization and standardization enable continuous health monitoring, issues such as missing data values and technical glitches can jeopardize data consistency and timeliness. On the other hand, the Efficiency in processing the large volume of data from disparate sources generated by the healthcare industry is critical. In this chapter, we try to provide an overview of how distributed computing and Artificial Intelligence can be used in the context of smart healthcare and big data in practical use cases, enabling insights to improve patient care. In addition, we propose a workflow for developing prognostic models that uses the SMART BEAR infrastructure and leverages the capabilities of the Big Data Analytics engine to standardize and harmonize data. Our workflow improves data quality by evaluating different imputation algorithms and selecting the one that preserves the distribution and correlation of features similar to the original data. We applied this workflow to a subset of data in the SMART BEAR repository and evaluated its impact on predicting future health conditions, such as cardiovascular disease and mild depression. We also explored the potential for model validation by clinicians in the SMART BEAR project, the transfer of subsequent actions within the decision support system, and the estimation of the required number of data points.

---

S. Maghool · V. Bellandi (✉) · P. Ceravolo  
Computer Science Department, Università Degli Studi di Milano, Via Celoria 18, Milano (MI),  
Italy  
e-mail: [Valerio.Bellandi@unimi.it](mailto:Valerio.Bellandi@unimi.it)

S. Maghool  
e-mail: [Samira.Maghool@unimi.it](mailto:Samira.Maghool@unimi.it)

P. Ceravolo  
e-mail: [Paolo.Ceravolo@unimi.it](mailto:Paolo.Ceravolo@unimi.it)

**Keywords** Internet of Things (IoT) · Smart healthcare · Machine learning · Analytics · Cloud computation

## Abbreviations

EHR	Electronic Health Record
AI	Artificial Intelligence
BDA	Big Data Analytics
CVD	Cardiovascular Disease
ML	Machine Learning
CNN	Convolutional Neural Networks
IoT	Internet of Things
HIPAA	Health Insurance Portability and Accountability Act
GDPR	General Data Protection Regulation
EU	European Union
MNAR	Missing Not At Random
MAR	Missing At Random
MCAR	Missing Completely At Random
NLP	Natural Language Processing
CI	Continuous Integration
CD	Continuous Delivery
CT	Continuous Training
CM	Continuous Monitoring
DSS	Decision Support Systems
FHIR	Fast Healthcare Interoperability Resources
ETL	Extract, Transform, Load
DAG	Directed Acyclic Graphs
RMSE	Root Mean Square Error
MAE	Mean Absolute Error
AE	Absolute Error

### 1.1 An Overview of the Current State of AI in Healthcare

In recent years, numerous efforts have been made to implement AI technologies in the healthcare domain. These technologies have been highly successful in assisting clinicians in various areas, from diagnosis by analyzing medical data, to pattern recognition, to assisting in finding appropriate treatments. It can also be used to support medical decisions by providing real-time assistance and insights to clinicians.

AI-based tools can improve accuracy, reduce costs and save time compared to traditional diagnostic methods. In addition, AI can reduce the risk of human error and provide more accurate results in a shorter time.

ML models for clinical trials are mostly based on supervised learning. In supervised learning, the model is trained against gold standards defined by a clinical expert, such as a chart review of 1000 patients with and without CVD, to identify the existing pattern between patients with CVD and those without. In this process, if some hypotheses are raised by clinicians, the researcher tries to prove these hypotheses using ML algorithms, but rejecting them needs more clinical and theoretical validation and evidence [8]. On the other hand, unsupervised ML models can also be used when there is a need to phenotype multiple conditions. These models are generally not as accurate as supervised models, but allow high throughput over a handful of thousands of variables with improved accuracy. In addition, unsupervised models have been used to build clinical models to predict disease progression, optimize diagnostics, and target treatment. In addition, unsupervised pattern recognition analyses identify subgroups of patient-patient similarity in a high-dimensional or graph-based space.

From the wide variety of current applications of AI in healthcare, we can recall:

- **Robotics** in order to provide high-precision surgical procedures.
- **Digital secretary** to provide effective intervention by alerting nurses by continuously monitoring the patient.
- **Machine Learning** to predict and analyze the patterns in medical data and facilitate decision-making by processing a large volume of data. ML can assist in managing workflow, and automate tasks in a timely and cost-effective manner.
- **Deep learning** by adding layers, for instance in CNN and data mining techniques, can help in identifying data patterns.
- **Image processing** to improve the finding of a large number of medical images and speed up the diagnosis stage.
- **Convert unstructured text** data, such as medical charts and prescriptions, to an easily readable format.
- **Statistical analysis** of patients' health records to evaluate the result of treatment.
- **Big data analysis** in order to provide personal recommendations by processing the historical data.
- **Predictive modeling** for prediction and prevention of possible risk or adverse event of treatments.

However, there are still many concerns regarding AI-based technologies in the health domain and their usage is not widely spread among clinicians. In order to fully take advantage of the potential of AI, various issues such as trust, explainability [16], and responsibility of these technologies should be discussed. Moreover, developed systems should consider taking into account fairness [4, 17] and guaranteeing the benefit for all [48].

The prerequisites of a fair and accurate AI system in the healthcare domain, are data quantity and quality. ML, as a subset of AI, uses data as resources in which the accuracy is highly dependent on these two features of the input data in order to overcome the challenges and complexity of traditional diagnosis procedure [40].

The healthcare sector is being profoundly impacted by the rapid expansion of digital data, the increasing computing power driven by advancements in hardware

technologies (such as graphics processing units), and the rapid progress of ML algorithms, particularly those based on deep learning.

Various companies are developing devices and services that can assist in improving user health by acquiring health information from daily life using a combination of IoT technologies and wearable devices.

Even though healthcare organizations can adopt AI systems and integrate them into existing workflows, successfully implementing predictive analytics requires high-quality data, appropriate infrastructure, and clinicians' oversight to ensure appropriate and effective interventions for patients.

## 1.2 Data in Healthcare AI

As a consequence of digitalization, like every other industry, the healthcare sector is also producing data at a high rate that presents many advantages and challenges simultaneously. This huge volume of data requires suitable infrastructure to manage and analyze in order to extract meaningful information. Electronic Health Record (EHR) systems facilitate the systematic digitized collection of patient data through electronic devices and information systems. The advantages associated with EHR adoption span both organizational and clinical realms [31]. Clinical decision support systems, computerized order entry systems, and health information exchange systems can significantly enhance their efficiency when integrated with EHRs [13]. This, in turn, leads to societal benefits such as decreased medical errors, enhanced research capabilities, and improved access to information for both patients and healthcare professionals [38]. Given the contemporary challenges that healthcare systems confront, particularly in the face of the rapid aging of populations [49], legislative bodies in the EU, the US, and other nations have responded by formulating recommendations and standards for healthcare organizations to follow in their utilization of EHRs to bolster operational efficiency [20, 52]. This trend coincides with a growing interest in mobile health (mHealth) monitoring systems. Alongside advancements in hospital infrastructure, mHealth is paving the road for the establishment of smart health ecosystems worldwide, leveraging data from mobile devices, wearables, and IoT devices both within and beyond healthcare facilities. These emerging smart health ecosystems enable the continuous collection of data from everyday life, which can be analyzed to obtain the evidence necessary to offer personalized interventions [7, 18].

The analysis of the extensive data gathered from EHR holds great promise in various healthcare applications. These applications include extracting clinically relevant information and facilitating diagnostic evaluations [33]. Additionally, EHR data can be used to generate real-time risk scores for patient transfer to intensive care [19]. It is also valuable in predicting in-hospital mortality, readmission risk, prolonged hospital stays, and discharge diagnoses [47]. Furthermore, EHR data aids in predicting future deterioration, such as acute kidney injury [51], and enhances decision-making strategies, including the weaning of mechanical ventilation [44] and the management

of sepsis [45]. Moreover, it assists in learning treatment policies from observational data [24].

Proof-of-concept studies have focused on streamlining clinical workflows. This involves tasks like automatically extracting semantic information from transcripts [28], recognizing speech during doctor-patient conversations [15], predicting the risk of patients failing to attend hospital appointments [41] and, summarizing doctor-patient consultations.

At the current status EHR systems have grown substantially in terms of data volume and diversity, posing a significant challenge in data management. Effective data management necessitates the formulation of a data strategy and the establishment of reliable methods for data storage, access, integration, cleansing, and data preparation for analytic [29].

### ***1.2.1 Big Data in Healthcare***

As is intuitively understandable from the term, “Big Data” refers to a large amount of heterogeneous data that is unmanageable using traditional software or internet-based platforms. It stores massive amounts of data generated from various sources such as IoT devices that create a continuous stream of data while monitoring the health of people (or patients) which makes these devices a major contributor to big data in healthcare. Such resources can interconnect various devices to provide a reliable, effective, and smart healthcare service to the elderly and patients. Big data encompasses both structured and unstructured data that organizations generate focusing on processing and analyzing data in its raw and unstructured form.

Therefore, we need advanced technological applications and software that can utilize fast and cost-efficient high-end computational power to make sense of this large amount of data. By leveraging distributed computing and parallel processing, big data platforms enable organizations to extract meaningful insights and patterns from massive datasets.

### ***1.2.2 Privacy and Security in Healthcare AI***

While the term “privacy” has been a frequent topic of discussion, yet a unified definition and clear procedure to ensure privacy-preserving policies remain elusive. This has led to ongoing confusion regarding the meaning, value, and scope of the concept of privacy. Privacy primarily revolves around the collection, storage, and utilization of personal information. It delves into questions like whether data should be collected in the first place and whether there are justifiable reasons for using data collected for one purpose in another (secondary) context. An essential aspect of privacy analysis pertains to whether individuals have given their authorization for specific uses of their personal information [9, 55].

Privacy plays a pivotal role in facilitating socially beneficial endeavors, such as health research. People are more inclined to take part in and endorse research when they have confidence in the protection of their privacy. Moreover, safeguarding privacy is regarded as a means to improve the quality of data for research and quality enhancement initiatives. When individuals take measures to safeguard their privacy, such as refraining from seeking healthcare or withholding information, it results in the introduction of inaccurate and incomplete data into the healthcare system. These flawed data, in turn, get used for research, public health reporting, and outcomes analysis, perpetuating the same vulnerabilities [22]. Medical records can include some of the most intimate details about a person's life. They document a patient's physical and mental health and can include information on social behaviors, personal relationships, and financial status [23].

Ensuring the security of data in health research is of utmost importance due to the substantial amount of personally identifiable health information collected, stored, and utilized in this field, much of which can be sensitive. In the event of a security breach, individuals whose health information has been inappropriately accessed may face a multitude of potential harms. Additionally, there is the risk of economic, social, and psychological harm.

Health Insurance Portability and Accountability Act<sup>1</sup> (HIPAA) is a U.S. federal law enacted in 1996 with the primary goal of protecting the privacy and security of individuals' health information, as well as ensuring the portability of health insurance coverage. On the other hand, the General Data Protection Regulation<sup>2</sup> (GDPR) is a comprehensive data protection regulation that became effective in the European Union (EU) in 2018. Its primary aim is to enhance the protection of individual's personal data and provide them with more control over how their data is used. GDPR applies not only to organizations within the EU but also to organizations outside the EU that process the personal data of EU residents. It covers a wide range of personal data, including health data. Adherence to both regulations is essential for protecting individual privacy and data security, while they have different scopes and areas of emphasis.

### ***1.2.3 Data Quality and Preprocessing***

The adoption of EHR systems necessitates the development of comprehensive data management procedures with a primary focus on "data quality" and "clinical significance". These two pillars are instrumental in harnessing data for enhanced monitoring and diagnostic procedures, particularly in the context of mobile health (mHealth) data. In mHealth, data records may be collected from various devices, at different times, and with varying levels of quality. Data transmission interruptions due to technical or usability issues, network problems affecting IoT device availability [32], tem-

---

<sup>1</sup> <https://www.cdc.gov/php/publications/topic/hipaa.html>.

<sup>2</sup> <https://gdpr-info.eu/>.



porary discontinuation of monitoring plans due to patient overload perceptions [54], misaligned time series arising from different temporal granularities in data collection [12], and gaps and missing values in time series [34] can all render time series incomplete, jeopardizing the validity of data analytics. The importance of verifying data completeness, consistency, and timeliness through tests and implementing methods to rectify or enhance data in cases of low quality is well-documented in the data quality literature [14]. Moreover, the effectiveness of prognostic analytics relies on the accuracy achieved by predictive models and the significance of the samples used for model training [42]. The design of a prognostic model encompasses considerations of domain complexity, domain stability, and sample size to achieve the required accuracy [6].

Whilst EHR systems are constantly producing and recording data, leveraging Machine Learning algorithms for creating prognostic and predictive models has implications for patients, caregivers, and healthcare facilities for cost management purposes [2]. Detecting systematically the issues concerning the data quality imposed by missing data is an interesting problem that is explored in [50]. The authors have found patterns in the condition domain and investigated the processes that shape them suggesting data quality issues influenced by system-wide factors that affect individual concept frequencies. The most general patterns identified in the literature are Missing Not At Random (MNAR), Missing At Random (MAR), and Missing Completely At Random (MCAR) [25]. MNAR points out that there is a relationship between the propensity of a value to be missed and its values. For example, people with the lowest education are not answering questionnaires including the questions on their educational courses. MAR refers to a category of missing values that are not related to other missing values but are related to observed values. For example, men are more likely to report their weight than women. MCAR, on the other hand, represents a category once there is no relationship between missing values and any other values. Nothing makes some data more likely to be missing than others. For example, blood pressure records are missing randomly, due to user ignorance or of charge battery. Even though in an approach the data scientists limit the study only to those patients with complete data, recent studies found that, compared to restricting the analysis, imputation techniques improved the accuracy of predictions at any proportion of missing data [26]. This implies that researchers should consider whether all the variables related to missingness can plausibly be included in the imputation model to limit bias and improve accuracy.

On the other hand, in the longitudinal studies, covered topics include reliability, validity, sampling, aggregation, and the correspondence between theory and method. More specifically, in these studies, practical issues in longitudinal research, such as the drop-out problem and issues of confidentiality are also addressed [11, 36], while the automation of this procedure is still missing. Moreover, due to the sensitivity of the health care domain, not only a deep knowledge of the process is needed but also continuous evaluation of the curation strategy should be considered.

### 1.3 Distributed Computing in AI

Distributed computing infrastructure refers to the hardware and software systems designed to support distributed computing, where multiple computers work together to solve complex problems or process large amounts of data. It involves the use of a network of interconnected computers, often referred to as nodes or servers, that collaborate and share resources to achieve a common goal [30]. Its major capabilities are balancing loads between computers that are taking care of computation while it is fault-tolerant and resilient, and requires robust monitoring and management tools to track the performance of distributed components.

The combination of Distributed computing and AI technologies allows for the efficient processing of large volumes of data that the healthcare industry generates from various sources and enables insights that can improve patient care, accelerate research, personalized medicine, disease prevention, healthcare cost reduction, and enhance healthcare operations. Distributed computing and AI can be utilized in the context of smart healthcare and big data in the following domains:

**Data Collection:** Distributed computing frameworks, such as Apache Hadoop<sup>3</sup> or Apache Spark<sup>4</sup>, can be employed to collect, store, and preprocess these diverse data types from multiple sources [1].

**Data Integration:** Distributed computing platforms facilitate the integration of disparate data sources by enabling parallel processing and data transformation. This integration allows healthcare professionals and researchers to combine and analyze data from different domains, uncovering correlations and patterns that were previously inaccessible.

**Data Preprocessing:** Big data in healthcare often needs preprocessing to clean, transform, and format the data for analysis. Distributed computing allows for parallel processing, making data preprocessing more efficient.

**Data Storage and Management:** Distributed databases, like HBase or Cassandra, can handle the storage and retrieval of vast amounts of healthcare data. This enables quick access to patient records and other critical information for real-time decision-making.

**Data Analysis and Machine Learning:** Distributed computing frameworks provide the computational power required to train complex models on massive datasets. These AI models can assist in diagnosing diseases, predicting outcomes, recommending treatments, and identifying potential risks or anomalies in real time.

**Real-Time Monitoring:** Distributed computing systems coupled with AI algorithms can enable real-time monitoring of patient data streams, including vital signs, activity levels, and medication adherence. By continuously analyzing this data, healthcare providers can identify critical events, detect early warning signs, and intervene promptly to prevent adverse outcomes.

---

<sup>3</sup> <https://hadoop.apache.org>.

<sup>4</sup> <https://spark.apache.org>.

**Predictive Analytics:** Big data analytics combined with distributed computing and AI can leverage historical data to develop predictive models. These models can forecast disease progression, identify high-risk patients, anticipate resource requirements, and support proactive interventions. Predictive analytics can aid in optimizing healthcare resource allocation and improving operational efficiency.

**Natural Language Processing (NLP):** NLP algorithms can extract valuable information from unstructured text in EHRs, medical literature, and patient feedback, providing valuable insights for healthcare professionals.

**Privacy and Security:** When working with sensitive patient data, privacy and security are of utmost importance. Distributed computing can incorporate privacy-preserving techniques such as secure multiparty computation and differential privacy to protect patient information while enabling collaborative analysis across multiple healthcare institutions [1].

**Scalability and Performance:** Big data in smart healthcare is characterized by its volume, velocity, and variety. Distributed computing architectures enable horizontal scalability, meaning the system can seamlessly add more computational resources as data volumes grow. By distributing the workload across multiple nodes, distributed computing improves performance and reduces processing times.

**Federated Learning:** Federated learning is an emerging approach that combines AI and distributed computing in healthcare. It enables the training of machine learning models on decentralized data sources while preserving data privacy. In this scenario, AI models are trained collaboratively across multiple healthcare institutions without sharing patient data, thereby fostering data privacy and security.

**Data Fusion:** In smart healthcare, data from various sources are fused together to provide a comprehensive view of the patient's health status. Distributed computing facilitates data fusion from diverse sources and enhances the overall quality of healthcare decision-making.

The integration of distributed computing and AI in smart healthcare allows for the efficient processing and analysis of big data. These technologies enable healthcare professionals to derive valuable insights, improve patient care, enhance operational efficiency, and advance medical research while ensuring data privacy and security.

## 1.4 Continuous Learning Models

ML algorithms learn from data without being explicitly programmed to do so. Traditional ML models are trained on the data assuming the future produced data distribution will be statistically more or less identical to the retrospective data, while it is not always a valid assumption and “concept drifts” are likely to occur to the dataset and to the model subsequently. This way, the model gets outdated and is not valid (able) anymore for prediction purposes.

In the following case scenarios adopting continuous learning is highly recommended: (I) To keep an ML model up to date on the latest data, once ML models fall below an acceptable threshold for the accuracy metrics; (II) Once the data can

become too statistically dissimilar from the data the ML model was originally trained on.

Learning from the prospective data contains four major components, *Continuous Integration (CI)*, *Continuous Delivery (CD)*, *Continuous Training (CT)*, and *Continuous Monitoring (CM)* to mimic a human's ability to acquire and fine-tune information continuously.

Using distributed computing infrastructures makes the CI of data more feasible than traditional approaches while CD helps that as new software features and fixes pass through the develop-build-test cycle, they become available as rapidly as possible.

Unlike traditional ML models, which are trained on a static dataset and require periodic retraining, continuous learning models iteratively update their parameters to reflect new distributions in the data, allowing them to remain updated and adaptive to the continuously changing data.

CM and model evaluation provide a means to assess the model's performance and reproducibility to ensure consistent outcomes during model evaluation.

Hence, for continuous learning from EHR systems, *data quality* and *clinical significance* become central workflows that these systems must address in the context of continuous data acquisition and model adaptation. Apriori evaluation of these dimensions is no longer aligned with the goals of modern EHR infrastructures, necessitating the design of software and data management workflows accordingly.

## 1.5 Clinical Decision Support

Explain how EHR data can be used to provide real-time clinical decision support to healthcare providers. Highlight the potential to reduce medical errors and improve patient care.

Envisioning the support received from computers in complex clinical situations by decision support systems (DSSs) that are designed to be used interactively by clinicians as they seek to reach decisions, regardless of the underlying analytic methodology that they incorporate, has taken place from a long time ago.

With the evolution of communication technologies and the digitalization of healthcare data, the ability to offer support to clinicians has faced a great improvement. For example, many built-in decision-support tools in medical devices create a variety of visualization and interpretation of medical data. Moreover, leveraging AI algorithms such as ML and NLP has broadened the clinical perspective by predictions using prospective data. DSS requires an interactive and intuitive design with a strong scientific foundation considering the established evidence for its safety, validity, and reproducibility of the results. The AI model used in DSS should be transparent while taking into account the relevance and the pertinent domain with which clinicians are likely to ask for assistance. After all, it should be noted that DSSs are designed for assistance, to reduce medical error, and to improve patient care by continuous monitoring using EHR and are not replaceable with a clinician.

## 1.6 A Real Case Scenario: The SMART BEAR Case Study

The SMART BEAR project<sup>5</sup> provides a comprehensive substructure for long-term continuous examinations and testing the well-being status of older people using wearable devices, mobile apps, and follow-up assessments by trained personnel and physicians.

SMART BEAR complements an EHR system by providing continuous monitoring, periodic assessments, data gathering from different resources, and providing both descriptive and predictive analyses.

Utilizing the medical/clinical data requires much effort in unifying the concepts and terms to make the data understandable and usable by other clinicians and scientists. A proposed solution is leveraging the unique LOINC<sup>6</sup> and SNOMED-CT<sup>7</sup> codes in defining observations, encounters, and biological considerations. Data storage in the SMART BEAR is designed based on the standardized data acquisition procedures specified in the *Mapping on Fast Healthcare Interoperability Resources (FHIR)* by [37]. FHIR provides a means for representing and sharing information among clinicians and organizations in a standard way regardless of the ways local EHRs represent or store the data to advance interoperability.

The data measured and collected with *SMART BEAR* devices, mobile applications, and questionnaires will be stored in HAPI FHIR repositories using the unified codes. Regarding the integration of questionnaires on the FHIR repository, a generic model is defined<sup>8</sup>.

In this chapter, we aim to present the results attained within the SMART BEAR project, focusing on the development of a comprehensive data management pipeline for continuous learning in EHR systems. Our solution incorporates multiple data management procedures into automated and modular workflows, enabling organizations to cultivate a culture of continuous improvement.

### 1.6.1 The SMART BEAR Infrastructure

The main objective of the SMART BEAR project is continuous and objective monitoring of *quality of life* of elderly people and their ability to live independently [18]. Considering these objectives, in order to implement efficient and valid analytics in a continuous data acquisition environment, it is required to cure the received data in multiple stages of the data management process [5].

Figure 1.1 presents the technical infrastructure of the SMART BEAR project [43]. The infrastructure contains different components such as Big Data Analysis (BDA) Engine, Security Component, Decision Support System (DSS), Dashboard, and Data

---

<sup>5</sup> <https://www.smart-bear.eu/>.

<sup>6</sup> <https://loinc.org/>.

<sup>7</sup> <http://www.snomed.org/snomed-ct/Use-SNOMED-CT>.

<sup>8</sup> <https://www.hl7.org/fhir/questionnaireresponse.html>.

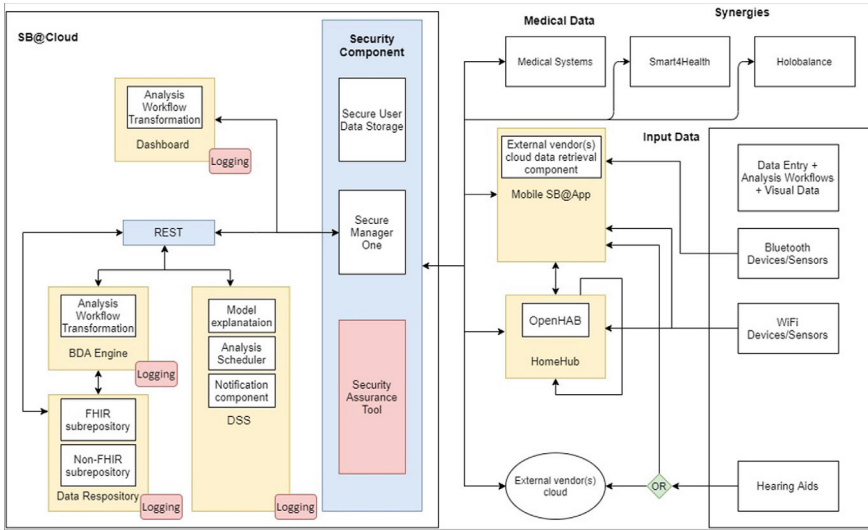


Fig. 1.1 An overview of the SMART BEAR infrastructure as presented in [43]

Repository. The received data from home sensors and other synergy studies are depicted on the right side of the picture. All the data curation procedures, including workflows related to data quality assessment, data preparation, sample size evaluation, and continuous learning, are supposed to take place in the BDA engine. Furthermore, the power of the BDA engine is exploited for prediction and personalized intervention purposes.

Taking advantage of the capabilities of the distributed systems, the BDA engine is tailored to provide scalability in terms of the addition of resources to the platform to support the increase in workload. Due to the execution on Docker containers, the configuration and deployment of resources are easy with high flexibility.

The BDA engine addresses the functionalities required for processing DAWs (Data Analysis Workflows) and storing execution results. It uses a series of suitably configured Open-Source components and a custom-developed one responsible for piloting the execution of the various analyses available in the catalog for results storage and presenting them in the dashboard.

1.6.1.1 BDA Engine Components

The components used by the platform are as follows:

- **Apache Hadoop:** From the Hadoop ecosystem, an exclusively distributed file system (HDFS) is utilized for storing the extracted data from the FHIR repository and transforming it in a tabular format appropriately.

- **Apache Hive Metastore**<sup>9</sup>: It contains all the information regarding the databases, the tables, and the relationships between them. This is especially useful for handling data that is transformed and saved on the distributed file system. The Metastore allows interoperability between the various components that need to access the data and enables to launch SQL-like queries using Trino and Spark.
- **Apache Spark**: This is the component that allows data management, data engineering, and machine learning tasks on a large dataset. Thanks to its abilities, the data are accessible through the Metastore considering the HDFS data similar to tables in a SQL database, simplifying life for those who have to develop the analytics and guaranteeing the necessary scalability for the platform.
- **Trino**<sup>10</sup>: It is a distributed SQL query engine designed to query large data sets distributed over one or more heterogeneous data sources. It is designed to handle data warehousing and analytics: data analysis, aggregating large amounts of data, and producing reports (OLAP). It could be leveraged for Extract, Transform, Load (ETL) transformations without the overhead of Spark. Furthermore, Trino with the proper configuration can run SQL queries directly on tables stored in HDFS.
- **Apache Airflow**<sup>11</sup>: It can develop, schedule, and monitor batch-oriented workflows. Airflow's extensible Python framework enables us to build workflows using different technologies, combined with docker, it can run custom images with all the tools needed to run analytic tasks.
- **BDA API**: It is internally developed for providing a REST-type interface to the dashboard and other platform components in order to be able to interact with workflows and save/retrieve the results of previous executions. The BDA API includes a catalog of atomic analytics that could be composed in workflows the engine schedule automatically or in dependence on specific events.
- **Delta Lake**<sup>12</sup>: It is an open-source storage framework that enables building a Lakehouse architecture. It is located on top of Hadoop providing ACID Transactions, and scalable metadata handling while unifying streaming and batch data processing on top of existing data lakes like HDFS .
- **Apache Zeppelin**<sup>13</sup>: It is a Web-based tool that enables users to create interactive data analysis, prototype some of the analytics that should be translated into a proper workflow, and share preliminary results in a collaborative environment for the data scientist.

---

<sup>9</sup> <https://hive.apache.org>.

<sup>10</sup> <https://trino.io>.

<sup>11</sup> <https://airflow.apache.org>.

<sup>12</sup> <https://delta.io>.

<sup>13</sup> <https://zeppelin.apache.org>.

### 1.6.1.2 Data Storing and Data Management by BDA Engine

Exported data from various storage sources, such as the FHIR repository, is transformed through a series of steps and stored in the format requested by Delta Lake. This format allows for building a Lakehouse architecture. The Lakehouse is an open architecture that combines the best elements of data lakes and data warehouses. As some of the key features of this kind of architecture, we can name: (i) ACID transactions, (ii) Schema enforcement, (iii) Business intelligence support, (iv) Openness, (v) Support for diverse workloads, and (vi) Support for structured and unstructured data.

Through Delta Lake, we support all these requirements and offer a real version of the management process for data in our environment. With the *Time Travel* feature it is possible to specify the version of a table. This allows us to be able to relaunch an analysis on a specific version of the data in order to be able to make a comparison between the results of different snapshots. It also provides a mechanism to replicate a result and audit the data and the obtained results.

### 1.6.1.3 Data Analysis Tasks

In the SMART BEAR project, data analysis tasks are mainly performed using Spark and Python libraries that can operate on data in Delta Lake format. Depending on the complexity of the workflows, these tasks could be composed of several steps. In some cases, it is easier to query and prepare data through Spark and then use other more specific libraries or tools to perform the required analytics. With Apache Airflow we can create DAGs (Directed Acyclic Graphs) composed of different tasks and use HDFS as a distributed file system to save the data needed for the various steps. Furthermore, with this flexibility, if needed, it is possible to use ready Docker images containing all the required tools for creating models and performing analysis. The main requirement therefore always remains to be able to read data from a distributed file system such as HDFS.

### 1.6.1.4 Data Flow Management and Workflow Orchestration

The most complex job in data flow management is querying the FHIR repository to export the data from this repository and insert it into the tables in Delta Lake. The workflow, that runs each day, is therefore composed of different tasks which can be roughly divided as follows: (i) Export from FHIR, (ii) Flattening the data, and (iii) Upsert in Delta Table. To export the data, the possibilities offered by the FHIR REST API in using the Bulk Data Export API, are exploited. Therefore it is also possible to request to export data that has been inserted/modified in the repository from the last time the export process was successful. The data exported in *ndjson* format are then stored on HDFS, the next step is to transform them into a flat type format as if they were data belonging to a common SQL table. All of this is orchestrated using



Apache Airflow as the main engine. Using Airflow, it is convenient to configure the workflows' execution according to predefined intervals. In very specific cases it is possible to execute workflows using schedules defined through a specific interval by a Cron-type expression.

### 1.6.1.5 Data Visualization

For data visualization purposes, the BDA Engine mainly makes use of two different tools: Apache Echarts<sup>14</sup> to display the results of the analytics in the Dashboard interface and Apache Zeppelin to allow data scientists to carry out data exploration and to create examples of analytics which will be deployed in production. Apache Echarts is an open-sourced JavaScript visualization tool that can run on web browsers and mobile devices, it also provides a rich library of basic charts and the possibility to extend or customize according to the needs of the asked output. There are many available features in this tool, among the more common ones are: (i) *Datazoom* that is used for zooming a specific area, which enables users to investigate data in detail, get an overview of the data, or get rid of outlier points. (ii) *Timeline* which provides functions like switching and playing between multiple charts with relative time differences. (iii) *Toolbox* that contains some functionalities such as the export to PNG format. (iv) *Legend* that shows symbol, color, and name of different series. The user can click legends to toggle displaying series in the chart.

### 1.6.1.6 Platform Management

Since all the tools used in the BDA Engine have been containerized, to manage the entire platform the tool we rely on is the deployment platform itself, i.e., Kubernetes (K8s). K8s is the system that we use to handle scaling, automatic system deployment, and manage all the containerized applications that we use in the Cloud. All the components described above have their basic configurations saved as ConfigMap inside the SMART BEAR repository dedicated to the BDA Engine while passwords and sensitive data are stored as Secrets. For some components, it was decided to use Helm Charts<sup>15</sup> to deploy as the Spark cluster, while for others ad-hoc deployments are used. Then the final configurations should be done using the WEB UI provided by the tool. These administration UIs are not available to all users, but only to system administrators who can also modify the deployment of the various tools and manage the resources needed to keep the infrastructure fast enough.

---

<sup>14</sup> <https://echarts.apache.org>.

<sup>15</sup> <https://helm.sh>.