Digital Health and Informatics Innovations for Sustainable Health Care Systems
J. Mantas et al. (Eds.)

doi:10.3233/SHTI240624

INSAFEDARE Project: Innovative Applications of Assessment and Assurance of Data and Synthetic Data for Regulatory Decision Support

Parisis GALLOS[a,1], Nicholas MATRAGKAS[b], Saif ul ISLAM[c], Gregory EPIPHANIOU[c], Scott HANSEN[d], Stuart HARRISON[e], Bram VAN DIJK[f], Marcel HAAS[f], Giorgos PAPPOUS[g], Simon BROUWER[h], Francesco TORLONTANO[i], Saadullah Farooq ABBASI[j], Omid POURNIK[j], James CHURM[j], John MANTAS[a], Carlos Luis PARRA-CALDERÓN[a], Dimitrios PETKOUSIS[a], Patrick WEBER[a], Benjamin DZINGINA[b], Chokri MRAIDHA[b], Carsten MAPLE[c], Jim ACHTERBERG[f], Marco SPRUIT[f], Evi SARATSIOTI[g], Younes MOUSTAGHFIR[h], and Theodoros N. ARVANITIS[j]

[a] *European Federation for Medical Informatics, Switzerland*
[b] *CEA, List, Université Paris-Saclay, France*
[c] *University of Warwick, UK*
[d] *The Open Group, UK*
[e] *ETHOS Digital Health, UK*
[f] *Leiden University Medical Center, The Netherlands*
[g] *Health Technology Certification Limited, Cyprus*
[h] *Syntho BV, The Netherlands*
[i] *Istituto Italiano Per La Privacy, Italy*
[j] *Department of Electronic, Electrical and Systems Engineering, School of Engineering, University of Birmingham, Birmingham, UK*

ORCiD ID: Parisis Gallos https://orcid.org/0000-0002-8630-7200, Nicholas Matragkas https://orcid.org/0000-0002-8594-1912, Saif ul Islam https://orcid.org/0000-0002-9546-4195, Gregory Epiphaniou https://orcid.org/0000-0003-1054-6368, Stuart Harrison https://orcid.org/0000-0003-3873-4512, Bram van Dijk https://orcid.org/0009-0002-9176-1608, Marcel Haas https://orcid.org/0000-0003-2581-8370, Simon Brouwer https://orcid.org/0000-0002-0956-0851, Saadullah Farooq Abbasi https://orcid.org/0000-0001-9814-3023, Omid Pournik https://orcid.org/0000-0001-7938-0269, James Churm https://orcid.org/0000-0003-0654-5960, John Mantas https://orcid.org/0000-0002-3051-1819, Carlos Luis Parra-Calderón https://orcid.org/0000-0003-2609-575X, Patrick Weber https://orcid.org/0000-0003-4469-0464, Chokri Mraidha https://orcid.org/0000-0003-2993-5734, Carsten Maple https://orcid.org/0000-0002-4715-212X, Jim Achterberg https://orcid.org/0009-0000-9589-7831, Marco Spruit https://orcid.org/0000-0002-9237-221X, Theodoros N. Arvanitis https://orcid.org/0000-0001-5473-135X

[1] Corresponding Author: Parisis Gallos, European Federation for Medical Informatics (EFMI), Ch de Maillefer 37, CH-1052 Le Mont-sur-Lausanne, Switzerland; E-mail: parisgallos@yahoo.com.

Abstract. Digital health solutions hold promise for enhancing healthcare delivery and patient outcomes, primarily driven by advancements such as machine learning, artificial intelligence, and data science, which enable the development of integrated care systems. Techniques for generating synthetic data from real datasets are highly advanced and continually evolving. This paper aims to present the INSAFEDARE project's ambition regarding medical devices' regulation and how real and synthetic data can be used to check if devices are safe and effective. The project will consist of three pillars: a) assurance of new state-of-the-art technologies and approaches (such as synthetic data), which will support the validation methods as part of regulatory decision-making; b) technical and scientific, focusing on data-based safety assurance, as well as discovery, integration and use of datasets, and use of machine learning approaches; and c) delivery to practice, through co-production involving relevant stakeholders, dissemination and sustainability of the project's outputs. Finally, INSAFEDARE will develop an open syllabus and training certification for health professionals focused on quality assurance.

Keywords. Medical Devices, Software as a Medical Device, Regulation, Synthetic Datasets

1. Introduction

Medical technology includes various products and services aimed to improve patient health. Medical devices, in vitro diagnostics, and digital health solutions are included in the medical technology domain. The medical technology sector in Europe is significant, providing numerous jobs and contributing substantially to healthcare expenditure [1]. Thanks to innovations like machine learning, AI, and data science, digital health is seen as promising for enhancing healthcare delivery and patient outcomes, enabling integrated care systems [2,3]. However, improper use of these technologies can have harmful consequences, prompting regulatory bodies to implement certifications for devices posing potential risks. Regulations like the Medical Devices Regulation (MDR) and the In Vitro Devices Regulation (IVDR) in the EU set device safety and effectiveness requirements. Software is also regulated as a medical device if it aids in diagnosis or treatment (Software as a Medical Device - SaMD). While traditional device assurance relies on predictability and determinism, emerging technologies like AI challenge this by operating in less predictable ways. These technologies learn from large datasets and apply this learning to individual patient data, providing diagnoses or recommendations with a degree of confidence rather than certainty [4,5]. Stakeholders need help to interpret regulatory frameworks for emerging technologies like machine learning. Recently, there has also been discussion on establishing a safety assurance framework for AI and machine learning in healthcare [6]. Efforts from regulatory bodies and industry groups aim to guide in navigating these challenges [7]. One proposed solution to validate digital health devices is using Realistic Synthetic Datasets (RSDs) [8,9]. These synthetic datasets mimic real data statistically [10-12], providing advantages such as bypassing approval processes for real data, including sensitive variables, and preventing cross-referencing with other datasets. Techniques for generating synthetic data from real datasets are highly advanced and continually evolving [13]. Initially, the field concentrated on using Classification and Regression Trees (CART) models [14] to create structured synthetic datasets. Later, it progressed to employing Generative Adversarial Networks (GANs) [15] for structured datasets with independent records. This paper aims to present the ambition of the INSAFEDARE project regarding the regulations of medical devices and how real and synthetic data can be used to assess the safety and effectiveness of these devices.

2. INSAFEDARE Objectives and Approach

INSAFEDARE scope is to provide a toolkit to enable cost-effective and high-assurance decision-making in the context of the processes that all stakeholders may follow as part of the regulation of medical devices. This toolkit includes scientific advice on ensuring quality, tools to find and check data, and a group where people can share advice for ongoing support. The project will examine how real and synthetic data can be used to check if devices are safe and effective. It will advise on how to ensure that datasets used for validation are good enough and how to validate devices using data-driven methods. INSAFEDARE will explore using made-up datasets to check if devices are safe before they go through the usual certification process. This could help developers lower their risks and prevent regulatory bodies from wasting time. Additionally, the project will share its findings publicly to create a standard for how this work should be done. It will also offer training to help people involved in regulation augment their knowledge base. The project will consist of three main pillars:

2.1. Assurance of large data-based validation

Assurance of new state-of-the-art technologies and approaches (such as synthetic data), which will support the validation methods as part of regulatory decision making. The project aims to assess various scenarios involving digital health interventions and innovations, examining how they rely on the use and reliability of data. It will conduct a thorough analysis to pinpoint any gaps in data assurance coverage within existing regulations. Additionally, the project will seek out standards, guidelines, and best practices concerning the safety and quality assurance of data-driven digital health innovations. This involves studying academic research and comparing regulatory strategies worldwide, including ongoing efforts to establish standards and advice on adapting existing regulatory frameworks to accommodate new technologies. In addition, another objective is the development of a safety and quality assurance framework for data in regulatory decision-making and clinical studies and to define a technology-forward hybrid real-world and synthetic data assurance framework in the context of current regulatory and legal obligations. In this sense, the application of FAIR principles by the design of data sets, with the enrichment of all data provenance information to allow human and computational (machine) access to data, will help the safe use of real data combined with synthetic data.

2.2. Research and Technology

This pillar of the project focuses on technical and scientific endeavours related to ensuring data safety and effectively utilising datasets. It emphasizes the use of machine learning techniques. Specifically, the project aims to explore machine learning methods for creating synthetic datasets that prioritise privacy. Additionally, it will delve into the evidence necessary to ensure the reliability of big data applications. This encompasses applications that utilise data to train predictive models, such as those employing machine learning, as well as applications utilising synthetic data generated through techniques like machine learning. Synthetic datasets must demonstrate equivalence to real data for their intended application, requiring statistical evidence to validate relevant attributes and advanced techniques such as machine learning-based classifiers to prove equivalent performance. The project will develop a trustworthy heritage query tool for integrating multiple datasets, considering the application of FAIR principles by design that will allow the joint analysis of data from different origins and of different natures (real and synthetic). Lastly, INSAFEDARE will analyse the compatibility of data-based approaches with privacy and open science principles and regulations.

2.3. Delivery to practice

The last pillar includes the delivery to practice through co-production involving relevant stakeholders, dissemination and sustainability of the project's outputs. The development and maintenance of a Co-Production based on publicly available guidance, data-driven assurance, and application validation will be held. Meanwhile, the project aims to develop a syllabus and training material for synthetic and real-world data-based assurance in regulation. In addition, INSAFEDARE will develop a digital web-based application, which will facilitate the regulatory processes, and offer support to regulators and manufacturers to manage changes during the entire life of an application. The tool will be in the form of a portal where manufacturers can access synthetic datasets that will serve as benchmarks, validating the manufacturer's results with an independent (common) dataset. The tool will be supported by an ontology coding the steps as well as information needed as part of the proposed regulatory framework. This way, manufacturers can share their information in a machine-readable format.

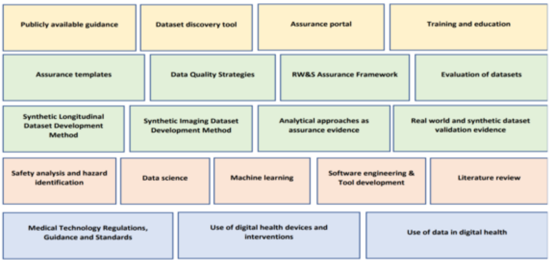
Publicly available guidance
Dataset discovery tool
Assurance portal
Training and education
Assurance templates
Data Quality Strategies
RW&S Assurance Framework
Evaluation of datasets
Synthetic Longitudinal Dataset Development Method
Synthetic Imaging Dataset Development Method
Analytical approaches as assurance evidence
Real world and synthetic dataset validation evidence
Safety analysis and hazard identification
Data science
Machine learning
Software engineering & Tool development
Literature review
Medical Technology Regulations, Guidance and Standards
Use of digital health devices and interventions
Use of data in digital health

Figure 1. Overview of the project's concept.

Figure 1 outlines the main components of the project. The blue blocks present the foundational domain knowledge, including examining medical technology regulations. The red blocks highlight the methodological approaches to be employed in the project. Green blocks show the creation of scientific and engineering advancements that will form the project's outcomes. Lastly, the yellow blocks provide an overview of the tools and results accessible to regulatory decision-makers. These include public guidance, tools for discovering datasets and generating synthetic datasets, an assurance portal, and a set of training materials and educational syllabi for stakeholders.

3. Expected Outcomes - Conclusions

INSAFEDARE will develop tools to help find, combine, and analyse different health datasets. At the same time, the project will develop guidance on validation and assurance of digital health applications and innovations (e.g., diagnosis, prediction, clinical decision support) using real-world and synthetic datasets. Synthetic data utility and privacy evaluation metrics will be implemented to stimulate standardisation and widespread acceptance of such metrics. In the project, a tool to track devices over time will be created so that any new evidence from new datasets can be considered. To this end, INSAFEDARE will develop training material on the assurance of datasets, good practice on dataset generation including guidance for organizations running clinical trials, and assurance of digital health applications. The training material will cover all stakeholder viewpoints and will consist of online short courses, as well as professional courses. Finally, INSAFEDARE will develop an open syllabus and training certification for health professionals on quality assurance.

Acknowledgements

This work is co-funded by the HORIZON.2.1 - Health Programme of the European Commission, Grant Agreement number: 101095661 - Innovative applications of assessment and assurance of data and synthetic data for regulatory decision support (INSAFEDARE).

References

[1] The European Medical Technology Industry in figures 2021, MedTech Europe.

[2] ASSESSING THE IMPACT OF DIGITAL TRANSFORMATION OF HEALTH SERVICES Report of the Expert Panel on effective ways of investing in Health (EXPH) https://ec.europa.eu/health/system/files/2019-11/022_digitaltransformation_en_0.pdf

[3] Global strategy on digital health 2020-2025 ISBN 978-92-4-002092-4 (electronic version) World Health Organization 2021 https://www.who.int/docs/default-source/documents/gs4dhdaa2a9f352b0445bafbc79ca799dce4d.pdf

[4] DCB0129: Clinical Risk Management: its Application in the Manufacture of Health IT Systems, https://digital.nhs.uk/data-and-information/informationstandards/information-standards-and-data-collections-including-extractions/publications-and-notifications/standards-and-collections/dcb0129-clinical-riskmanagement-its-application-in-the-manufacture-of-health-it-systems.

[5] ISO 81001-1:2021, https://www.iso.org/standard/71538.htm

[6] FDA, Artificial Intelligence and Machine Learning in Software as a Medical Device, https://www.fda.gov/medical-devices/software-medical-devicesamd/artificial-intelligence-and-machine-learning-software-medical-device

[7] Bellovin SM, Dutta PK, Reitinger N. Privacy and synthetic datasets. Stan. Tech. L. Rev.. 2019;22:1. Available at SSRN: https://ssrn.com/abstract=3255766 or http://dx.doi.org/10.2139/ssrn.3255766

[8] Buczak AL, et al. Data-driven approach for creating synthetic electronic medical records. BMC medical informatics and decision making. 2010 Dec;10:1-28.. https://doi.org/10.1186/1472-6947-10-59

[9] Moniz L, et al. Construction and validation of synthetic electronic medical records. Online J Public Health Inform. 2009;1(1):ojphi.v1i1.2720. doi: 10.5210/ojphi.v1i1.2720. Epub 2009 Dec 10.

[10] Baowaly MK, Liu CL, Chen KT. Realistic data synthesis using enhanced generative adversarial networks. IEEE 2nd Inter Conf on AI and Knowledge Engineering (AIKE) 2019 Jun 3 (pp. 289-292). IEEE..

[11] McLachlan S, et al. The ATEN framework for creating the realistic synthetic electronic health record. 11th Inter Confer on Biomed Engin Syst and Techn (BIOSTEC) – Vol 5: HEALTHINF. 2018: 220-30.

[12] Schiff S, Gehrke M, Möller R. Efficient enriching of synthesized relational patient data with time series data. Procedia Computer Science. 2018 Jan 1;141:531-8.

[13] Lin Z, Jain A, Wang C, Fanti G, Sekar V. Generating high-fidelity, synthetic time series datasets with doppelganger. arXiv preprint arXiv:1909.13403. 2019.

[14] Nowok B, Raab GM, Dibben C. synthpop: Bespoke creation of synthetic data in R. Journal of statistical software. 2016 Oct 28;74:1-26.

[15] Arvanitis TN, et al. A method for machine learning generation of realistic synthetic datasets for validating healthcare applications. Health Informatics Journal. 2022 Feb 12;28(2):14604582221077000.